

Neural Network Support Vector Detection via a Soft-Label, Hybrid K-Means Classifier

Robert A. Murphy, Ph.D.

Email: robert.a.murphy@wustl.edu

Abstract—We use random geometric graphs to describe clusters of higher dimensional data points which are bijectively mapped to a (possibly) lower dimensional space where an equivalent random cluster model is used to calculate the expected number of modes to be found when separating the data of a multi-modal data set into distinct clusters. Furthermore, as a function of the expected number of modes and the number of data points in the sample, an upper bound on a given distance measure is found such that data points have the greatest correlation if their mutual distances from a common center is less than or equal to the calculated bound. Anomalies are exposed, which lie outside of the union of all regularized clusters of data points.

Similar to finding a hyperplane which can be shifted along its normal to expose the maximal distance between binary classes, it is shown that the union of regularized clusters can be used to define a hyperplane which can be shifted by a certain amount to separate the data into binary classes and that the shifted hyperplane defines the activation function for a two-class discriminating neural network. Lastly, this neural network is used to detect the set of support vectors which determines the maximally-separating region between the binary classes.

I. INTRODUCTION AND RELATED WORKS

Inherent in traditional support vector detection is the assumption that the data are separable into two distinct classes, with the data points in each class being labeled distinctly, while minimizing mislabeling instances. In the case of linear support vector detection, a single hyperplane is sought, whereby the single hyperplane can be optimally shifted along its normal vector and its reflection. The shifted hyperplane and its reflection across the original, single hyperplane defines an optimal intermediate region, along with two distinctly labeled classes.

Assume the existence of a training set with known labels derived from a set of two distinct values. Since the weight vector for the single hyperplane defines its normal vector, then the problem reduces to finding a single weight vector such that its squared norm is minimized, subject to the constraint that the individually labeled training examples lie to one side of the original hyperplane, along with others sharing the same label. After which, by combining both the squared norm term and the equation for the original hyperplane, the constrained optimization problem is converted to an unconstrained optimization problem and solved using the method of Lagrange multipliers. The reader is referred to [1, pp. 311 – 314] for mathematical formalities on linear support vector machines.

In case the data are not linearly separable, the aforementioned algorithm will not work, as stated. *Slack variables* have to be introduced, which has the effect of reducing the area of the intermediate region, resulting in the constraint being modified. Furthermore, in the unconstrained problem, the slack

variables are taken into account by adding a penalty term as something on the order of the number of misclassifications encountered, where misclassifications take one of two forms. Either some data points are counted as *unclassified*, because they cannot be separated from the intermediate region, or they are misclassified, because they are incorrectly seen as training examples which belong to the class containing data points with a different label. The reader is referred to [1, pp. 315 – 318] for mathematical formalities on non-linear support vector machines.

For data sets which are extremely large, relative to its number of attributes, Mangasarian, et.al. showed in [55] that a linear support vector machine suffices when separating the data into two distinct sets with small error. Yet, in other kinds of problems, where the data are not linearly separable, or where a linear support vector machine gives rise to a significant number of misclassifications or an intermediate region with a high number of unclassified data points, the requirement for a non-linear support vector machine can be too expensive, both in time complexity and memory usage [17].

The prevailing methods for dealing with the expense associated with solving the non-linear optimization problem are variations of the *decomposition* algorithm, [40], [42], [45], [63], whereby the data are projected to a lower dimensional subspace, whose basis vectors are chosen such that the projected data are linearly separable. By linear separability of the data, the unconstrained, non-linear problem can be solved in the subspace via a linear support vector machine, resulting in a sequence of solvable sub-problems in each projected subspace. Influenced by Keerthi, et.al [44], Chung, et.al. [17] propose the use of specific seed values for the set of Lagrange multipliers of the linear support vector machines in each subspace such that faster convergence to the solution of the non-linear optimization problem is achieved. Likewise, the *chunking* method of Vapnik [75] makes use of the fact that the rows and columns of the matrix in the quadratic constraint, which correspond to Lagrange multipliers with a zero value, can be removed, while retaining the same solution through each iteration of a sub-problem.

Motivated to produce a method of solution which does not require iterative solutions of quadratic programs and does not require checking of certain conditions to ensure convergence to a solution, a method is provided such that the data are projected into a two dimensional space, whereby the projected data are either linearly separable or the data forms one contiguous class. When the data are separable, linear separation is achieved by forming a linear regression hyperplane through one of the classes and shifting it an appropriate amount in order to achieve separation of the classes. The associated shift

corresponds to the identification of the set of support vectors by measuring distances to the regression hyperplane from data points in the set which is in the complement of the set of data points used to form the regression hyperplane.

Suppose infinitely many copies of a bounded structure are used to partition \mathbb{R}^2 and let $\mathcal{B} \subset \mathbb{R}^2$ be a bounded subset containing finitely many copies of the bounded structure. Further suppose that structures in the partition are neighbors, if their respective boundaries have non-empty intersection and infinitely many of the bounded structures in the partition are individually occupied by exactly one point at the center of the structure, independently of all other structures. In [32] and [33], it is shown that, if the probability of neighboring structures each containing related points is greater than some critical value, then with probability 1, a path can be traced from any starting occupied bounded structure to any ending occupied bounded structure, with the path in between the start and the end consisting entirely of neighboring occupied structures. From this statement, the contrapositive statement is obtained such that, if the probability of neighboring structures each containing related points is less than or equal to the same critical value, then with probability 1, no such path exists for any two bounded structures. Hence, all points are either related to no other points or only related to finitely many points in neighboring bounded structures. It is this contrapositive condition that is of interest and great use during K -means classification, as classes are formed by groupings of inter-related data points.

Rarely does inter-related, real-world data conform to a predefined, rigid partition, as described above. As such, after removing the rigid partition of \mathbb{R}^2 , suppose that the data are modeled by a node process which randomly generates points within \mathcal{B} according to some predetermined probability distribution. Points in \mathcal{B} are inter-related, if they are within a certain distance of one another or some common point, such as the average of a set of previously-grouped, inter-related points. In [57], it is shown that, with probability 1, there is a path of inter-related points between any two points in \mathcal{B} , if the number of points relative to the area of \mathcal{B} is beyond a certain critical number or, if the maximum distance between inter-related points is larger than a certain critical number.

In [60], it is proven in cor. (7.4.39) that an ordered set of data, which is assumed to be spatially uniformly distributed, will form clusters, i.e. classes, if the measured distances between data points (or some common data point in each class) are below a certain threshold, which is computed as a function of the number of data points sampled from the total population of data. Now, most datasets would be too computationally expensive to order. So, we can get around this limitation by making certain assumptions.

We can make the assumption that the node process generates points according to the normal distribution by making use of a theorem from [38] in probability called the Central Limit Theorem. In essence, this theorem states that any set of randomly distributed data with finite mean and variance will tend to be normally distributed as the sample size grows large. This accounts for the ubiquity of the normal distribution in nature and why it's safe to make an assumption of normality in most cases. As such, the node process is allowed to run until $J = M^2$ points are generated, as represented by a sequence

of independent, identically distributed random variables, X_1, X_2, \dots, X_J , each with mean 0 and variance 1. An order statistic is applied to these random variables so that $X_{k_1} < X_{k_2} < \dots < X_{k_J}$, where $\sigma(i) = k_i$ for $i \in \{1, 2, \dots, J\}$. Note that, by default, $\{X_{k_i}\}_{i=1}^J$ is a sequence of *dependent* random variables since for each $i \in \{2, 3, \dots, J\}$, the random variable X_{k_i} depends upon X_{k_j} for all $j < i$. Now, $\{X_{k_i}\}_{i=1}^J$ is a (less computationally expensive) ordering of a set of points generated by the node process. The edge space of the higher dimensional data is then embedded within the 2-dimensional plane in order to aid in the calculation of a threshold on the distance measure.

To make the assumption of uniformity of the ordered set of points and to perform clustering therein, it is first noted that the Beta distribution is the probability distribution of an order statistic of normally distributed random variables [54]. The Beta distribution shape parameters are then defined to be, $\alpha = 1 = \beta$, since the data points are injectively generated into exactly 1 structure of the uniformly partitioned, 2-dimensional, bounded region. Hence, the ordering of the sample can be assumed to be approximately uniform.

Finally, with the defined partition, it is shown that under certain conditions, no approximation of probabilities in the continuum is required to prove the existence of a path of any order, as in [57]. Instead, the probability of a long range path in the continuum is equivalent to the probability of a long range path, over the same set of points, in the presence of the defined posterior partition when the maximum radial distance between connected points falls within a certain bounded interval. On this bounded interval, the probability of the existence of a long range path rises sharply when points connect at distances within the bounded interval. Lastly, the probability measure in question is found to be a unique random cluster measure which realizes a set of conditional probability measures. As such, the node process samples from the collection of conditional probability measures to form classes, when points connect at a distance less than or equal to the critical length.

In [13], Cai, et.al. investigate the problem of partial connectivity of randomly distributed points in a bounded region by making the assumption that, relative to the size of the bounded region, the number of points to be generated is relatively small. As such, a Poisson-distributed node process generates an independent set of points in the designated region. Copies of a hexagon of some fixed, immutable size, which is not dependent upon distances between generated points, are used to partition the bounded region. Points in the region are deemed to *connect* to form an *open* edge, if after the region is partitioned, the points lie within the same hexagon or neighboring hexagons, where hexagons are *neighbors*, if their respective boundaries have non-empty intersection. Otherwise, the edge between two points is *closed*. Likewise, they define the logical points at the centers of two neighboring hexagons to *connect* to form an *open* edge, if each neighboring hexagon independently contains at least one of the generated points. In [13], Cai, et.al. compute probabilities as a function of the density of hexagons which are occupied by at least one point. They showed that if the number of hexagons in the fixed partition is unbounded and the number of points generated within the continuum of the bounded region is below (or at) a critical threshold, then the

probability of a majority of the occupied hexagons (and points contained therein) in the bounded region being connected in a contiguous path will tend to zero. On the other hand, if the density of occupied hexagons is within a short interval around the critical threshold, then a connecting path of hexagons or points, from any start to any end, occurs with probability that rises sharply from some small positive value to a value close to 1 for densities that fall within the range of the short interval.

[31, *Thm. (1.1)*] gives an estimate of the length of the short interval. If the area of the bounded region is assumed to be one, without loss of generality, then the estimate of the length of the short interval can never be any better than $\Theta\left(\log^{1/4}(n)\sqrt{\frac{\log(n)}{n}}\right)$, where n is the number of points generated within the bounded region by the node process.

This work uses the distance notion of connectivity without the presence of a partition, the same as in [13]. However, it markedly differs in that the prototypical hexagon used in the defined posterior partition of the bounded region is allowed to change in size, if the node process is stopped and started again after a partition has already been defined. Since the prototypical hexagon is allowed to change in size so that the logical centers are closer to (or further away from) each other and point density is inversely proportional to the maximum connection length, therein lies an added advantage when calculating the probability of a connected path of hexagons or points. Moreover, if the prototypical hexagon always shrinks as the node process generates more points, then it should be expected that the critical threshold and the length of the short interval around the critical threshold are intertwined. It is shown that this is, in fact, the case.

The rest of this work is organized as follows. In section (II), we formulate the notion of connectivity of randomly-generated data points in the continuum using random geometric graphs and prove certain continuity results of the probability measure of a class of graph properties. We use the continuity results to show the existence of a critical connectivity radius (equivalent density of data points) and prove that the length of the sharp threshold interval containing the critical radius is of a certain size, depending upon the number of data points and the length of the critical radius.

In section (III), we uniformly partition the bounded region into shapes of the same size and formulate the notion of connectivity of randomly-generated data points using the random cluster model. The continuity results of section (II) still hold true and are used to show the existence of a critical connectivity radius, with the associated sharp threshold interval length being of a certain size, depending upon the number of data points and the length of the critical radius. In addition, relationships between the graph properties and probabilities of the graph properties are proven, along with a result about the relationship between the critical radii. These results, as well as other results from the random cluster model, are used to give a practical estimate of the length of the sharp threshold interval and a lower bound estimate of the change in the probability of the class of graph properties. Finally, it is shown that under certain conditions which are best for K -means classification, the probabilities of the graph properties are equivalent and the critical radii are of the same length under both formulations.

In section (IV), given a fixed, ordered sample from an unknown group of multi-modal normal distributions, we estimate the mean number of clusters to form in a typical K -means classification. Likewise, we also estimate the length of the sharp threshold interval as a function of the expected number of classes and an estimate of the size of the critical connectivity radius is given as a function of the number of data points in the sample and the expected number of classes to form.

In section (V), using the setup and results from section (IV), we define the regularized cluster and the set of anomalies, then prove that the regularized cluster and the set of anomalies are linearly separable and show that the linear regression hyperplane defined by the regularized cluster can be shifted by a certain distance along its normal vector to separate the regularized cluster from the set of anomalies.

Finally, in section (VI), we show that the set of anomalies is a superset for the set of support vectors defining the shifted hyperplane which delineates the maximally separating region between the binary classes and that the resulting neural network detects all of the support vectors.

II. RANDOM GEOMETRIC GRAPHS

A. Definitions

Definition 1: A node process is a mapping $X : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that for any subset $\mathcal{B} \subset \mathbb{R}^2$, there is an $n \in \mathbb{N}$ and a subset $\mathcal{X}_n = \{x_k\}_{1 \leq k \leq n} \subset \mathcal{B}$ with $X(\mathcal{B}) = \mathcal{X}_n$.

Definition 2: Suppose $\mathcal{B} \subset \mathbb{R}^2$ and X is a node process that randomly generates independent points $\mathcal{X}_n = \{x_k\}_{1 \leq k \leq n} \subset \mathcal{B}$ according to some probability distribution. Let $d : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a distance measure. Points $x, y \in \mathcal{X}_n$ are said to be r -connected and form an r -open edge, if $d(x, y) \leq r$, for some fixed $r > 0$. Points $x, y \in \mathcal{X}_n$ are r -disconnected and form an r -closed edge otherwise.

Definition 3: Let E be the set of edges between points in \mathcal{X}_n . $G(\mathcal{X}_n; r)$ is defined to be the r -graph of the set of all r -open and r -closed edges from E between points in \mathcal{X}_n .

Definition 4: Given points $x, y \in \mathcal{X}_n$ and some fixed $r > 0$, denote the r -edge between x and y as $\langle x, y \rangle_r$. A subset of points $C \subseteq \mathcal{X}_n$ forms an r -connected cluster if and only if given any $x, y \in C$, there exists r -open edges $\langle x, a_1 \rangle_r, \langle a_1, a_2 \rangle_r, \dots, \langle a_{k-1}, y \rangle_r \in E$ connecting x to y , for points $\{a_1, a_2, \dots, a_{k-1}\} \subseteq C$.

Definition 5: Let \mathcal{A} be a set of graphs of E and $G(\mathcal{X}_n; r) \in \mathcal{A}$. \mathcal{A} is said to be an increasing property if and only if for $r' \neq r$ such that $G(\mathcal{X}_n; r) \subseteq G(\mathcal{X}_n; r')$, it is true that $G(\mathcal{X}_n; r') \in \mathcal{A}$.

Definition 6: Let Ω be the set of values taken by the random variables \mathcal{X}_n , with \mathcal{F} being any σ -algebra of subsets of Ω and P a probability measure on (Ω, \mathcal{F}) . If \mathcal{A} is a monotone (increasing) property and $\epsilon \in (0, \frac{1}{2})$, define

$$r(n, \epsilon) = \inf\{r > 0 : P(G(\mathcal{X}_n; r) \in \mathcal{A}) \geq \epsilon\}$$

and

$$\Delta(n, \epsilon) = r(n, 1 - \epsilon) - r(n, \epsilon).$$

If $\Delta(n, \epsilon) = o(1)$, then \mathcal{A} has a sharp threshold.

B. An Important Result

Theorem 7: [31, Thm. (1.1)] For increasing properties \mathcal{A} consisting of graphs of points $\mathcal{X}_n \subset \mathbb{R}^2$,

$$\Delta(n, \epsilon) = \Theta(r_c \log^{1/4}(n))$$

where

$$r_c = O\left(\sqrt{\frac{\log n}{n}}\right).$$

These writings will be concerned, at least in part, with estimating the length of this critical interval for a particular property \mathcal{A} , using this framework.

C. Procedure

Let $\mathcal{B} \subset \mathbb{R}^2$ be a bounded region containing the origin $\hat{0} = (0, 0)$ and let X be a node process such that $X(\mathcal{B}) = \mathcal{X}_n$ is a set of n points which are uniformly distributed spatially throughout \mathcal{B} , where n is a Poisson random variable which takes a particular value (denoted as n) with density parameter $\lambda = \lambda(n)$ indicating the expected number of points generated per unit area of \mathcal{B} . For some fixed $r > 0$, points in \mathcal{X}_n will connect if their mutual distance is within r . For fixed $\rho \in (\frac{1}{2}, 1)$, define $\mathcal{A}_{[n, \rho]}^r$ to be the set of all r -connected graphs of subsets of \mathcal{X}_n containing at least $100\rho\%$ of all generated points which contain $\hat{0}$.

Let $\epsilon > 0$ be given and let $r(n, \rho, \epsilon)$ be the least connectivity radius $r > 0$ such that $P(\mathcal{A}_{[n, \rho]}^r) \geq \epsilon$. It will be shown that $P(\mathcal{A}_{[n, \rho]}^r)$ is an increasing function of the connection radius r . The aim is to estimate the length of the interval of connectivity radii such that the occurrence of $\mathcal{A}_{[n, \rho]}^r$ increases in probability from a value of ϵ to a value of $1 - \epsilon$ on the interval of radii. On this interval will be associated a particular radius such that the probability of the occurrence of $\mathcal{A}_{[n, \rho]}^r$ is $1/2$. On the left half of the interval, it is more likely that classes will form, with each containing less than half of all points so that no one class contains the majority of data points. No one class containing the majority of data points is important since, in this event, no one class will contain all data points with probability 1, which is guaranteed by [57] in the continuum and [33] in the partitioned continuum.

As an integral step in estimating the length of the interval of radii, continuity in r and ρ of $P(\mathcal{A}_{[n, \rho]}^r)$ will be shown. As such, by continuity in ρ , for small $\delta > 0$, the probability of $\mathcal{A}_{[n, \rho]}^r$ is close to the probabilities of $\mathcal{A}_{[n, \rho + \delta]}^r$ and $\mathcal{A}_{[n, \rho - \delta]}^r$. Furthermore, by continuity and the increasing nature of $P(\mathcal{A}_{[n, \rho]}^r)$ in r , there exists $r_0 = r_0(n, \rho, \epsilon)$ such that $P(\mathcal{A}_{[n, \rho]}^{r_0}) = 1/2$. This particular radius of connectivity demarcates the point, beyond which, the generated set of points will transition from a set of disjoint classes to one giant, inter-related class of points, almost surely. Furthermore, for $\epsilon > 0$, this radius of connectivity is the center of the estimated interval of radii, upon which, $P(\mathcal{A}_{[n, \rho]}^r)$ increases from ϵ to $1 - \epsilon$.

D. Definitions

Definition 8: Given a fixed point, $y \in \mathcal{X}_n$, an r -connected component containing y is the subset of points $\langle C_y \rangle_r \subseteq \mathcal{X}_n$

containing y and every $x \in \mathcal{X}_n \setminus \{y\}$ having a set of r -open edges connecting x to y .

Definition 9: Given an r -open edge, $e = \langle x, y \rangle_r \in G(\mathcal{X}_n; r)$, an r -connected component containing e is the subset of points $\langle C_e \rangle_r \subseteq \mathcal{X}_n$ containing x and y together with every $z \in \mathcal{X}_n \setminus \{x, y\}$ having a set of r -open edges connecting z to both x and y .

Definition 10: Let \mathcal{E} be any σ -algebra of subsets of E such that $\emptyset, E \in \mathcal{E}$, any $A \in \mathcal{E}$ implies $A^c \in \mathcal{E}$ and all countable unions of subsets of \mathcal{E} is again in \mathcal{E} . Suppose $\{\eta_k\}_{k \geq 1}$ is a sequence of random variables on E taking values in \mathbb{R} . It will be said that η_k converges weakly to a random variable $\eta : E \rightarrow \mathbb{R}$ (written $\eta_k \Rightarrow \eta$), if

$$\begin{aligned} \lim_{k \rightarrow \infty} F_k(x) &= \lim_{k \rightarrow \infty} P(\eta_k \leq x) \\ &= P(\eta \leq x) \\ &= F(x) \end{aligned}$$

for all $x \in \mathbb{R}$.

E. The Event

1) *Bounded Number of Nodes:* Let $\langle C \rangle_r \subseteq \mathcal{X}_n$ be an r -connected component containing $\hat{0}$ such that $|\langle C \rangle_r| = N$ and define $\rho_n(C) = \frac{N}{n}$. For $\rho \in (\frac{1}{2}, 1)$, define the graph property of all connected components containing at least $100\rho\%$ of all available points by

$$\mathcal{A}_{[n, \rho]}^r = \{\langle C \rangle_r \subseteq \mathcal{X}_n : \rho_n(C) \geq \rho\}. \quad (1)$$

As in [31], for $\epsilon \in (0, \frac{1}{2})$, define

$$r(n, \rho, \epsilon) = \inf\{r > 0 : P(\mathcal{A}_{[n, \rho]}^r) \geq \epsilon\} \quad (2)$$

to be the critical radius at which $\mathcal{A}_{[n, \rho]}^r$ occurs with probability at least ϵ and define

$$\Delta(n, \rho, \epsilon) = r(n, \rho, 1 - \epsilon) - r(n, \rho, \epsilon) \quad (3)$$

to be the length of the continuum of radii upon which $\mathcal{A}_{[n, \rho]}^r$ increases in probability of occurrence from $\epsilon > 0$ to $1 - \epsilon > 0$.

2) *Unbounded Number of Nodes:* In the case of n being unbounded, define the corresponding graph property to be

$$\mathcal{A}^r = \{\langle C \rangle_r \subseteq \mathcal{X}_\infty : |\langle C \rangle_r| = \infty\}. \quad (4)$$

Define

$$r(\epsilon) = \inf\{r > 0 : P(\mathcal{A}^r) \geq \epsilon\} \quad (5)$$

to be the critical radius at which \mathcal{A}^r occurs with probability at least ϵ and define

$$\Delta(\epsilon) = r(1 - \epsilon) - r(\epsilon) \quad (6)$$

to be the length of the continuum of radii upon which \mathcal{A}^r increases in probability of occurrence from $\epsilon > 0$ to $1 - \epsilon > 0$.

F. Continuity Results

In order to prove the existence of $r_0 > 0$ such that $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1/2$, it will be shown that $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r . By properties of probabilities measures, $P(\mathcal{A}_{[n,\rho]}^r) \in [0, 1]$ and by prop. (76), it is true that $P(\mathcal{A}_{[n,\rho]}^r)$ is non-decreasing as a function of increasing $r > 0$. By thm. (7), it is true that $P(\mathcal{A}_{[n,\rho]}^r)$ increases from $\epsilon > 0$ to $1 - \epsilon > 0$ for fixed $\epsilon \in (0, \frac{1}{2})$. Then, by continuity, there exists $r_0 > 0$ such that $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1/2$. If I is any continuum of radii and $P(\mathcal{A}_{[n,\rho]}^I)$ is defined to be the set $\{P(\mathcal{A}_{[n,\rho]}^r) : r \in I\}$, then it is easily seen that r_0 is in the interior of any compact interval of radii I_ϵ such that $P(\mathcal{A}_{[n,\rho]}^{I_\epsilon}) = [\epsilon, 1 - \epsilon]$. Seeking a contradiction, suppose r_0 is in the boundary of I_ϵ . Since I_ϵ is compact and $P(\mathcal{A}_{[n,\rho]}^r)$ is continuous in r , then $P(\mathcal{A}_{[n,\rho]}^{r_0}) = \epsilon$ or $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1 - \epsilon$. Therefore, $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1/2$ implies $\epsilon = 1/2$. This is a contradiction since $\epsilon \in (0, \frac{1}{2})$. Thus, r_0 is in the interior of I_ϵ . Q.E.D.

Now, if it can be shown that r_0 is independent of ϵ , then $r_0 \in I_\epsilon$ for all $\epsilon \in (0, \frac{1}{2})$. Note that $r_0 \in I = \bigcap_k I_{\epsilon_k}$ for any sequence $\epsilon_k \rightarrow 1/2$. Clearly I is compact so that r_0 is in the interior of I . Therefore, either I is an interval or $I = \{r_0\}$. Suppose I is an interval of radii. Since r_0 is in the interior of I , then there exists $r'_0 < r_0 \in I$. Now, since $\epsilon_k \rightarrow 1/2$, then $P(\mathcal{A}_{[n,\rho]}^{r'_0}) = 1/2$ and $r'_0 < r_0 = \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) = \frac{1}{2}\}$. This is a contradiction. Therefore, $I = \{r_0\}$ so that r_0 is unique. Q.E.D.

Continuity of $P(\mathcal{A}^r)$ in r is proven in [57] and can be used for proving continuity of $P(\mathcal{A}_{[n,\rho]}^r)$ in r as follows. Let $\partial\mathcal{B}$ denote the boundary of \mathcal{B} and define $\mathcal{A}_{\mathcal{B}}^r = \{\hat{0} \leftrightarrow \partial\mathcal{B}\}$ to be the property that there is an r -connected cluster containing $\hat{0}$ and a point in $\partial\mathcal{B}$. By arguments in [57], continuity of $P(\mathcal{A}^r)$ in r is equivalent to continuity of $P(\mathcal{A}_{\mathcal{B}}^r)$ in r for all bounded regions \mathcal{B} containing $\hat{0}$. Clearly, $P(\mathcal{A}_{\mathcal{B}}^r) = P(\mathcal{A}_{\mathcal{B}}^r - \mathcal{A}_{[n,\rho]}^r) + P(\mathcal{A}_{\mathcal{B}}^r \cap \mathcal{A}_{[n,\rho]}^r)$ so that continuity of $P(\mathcal{A}_{\mathcal{B}}^r)$ in r implies continuity of $P(\mathcal{A}_{\mathcal{B}}^r \cap \mathcal{A}_{[n,\rho]}^r)$ in r . Now, there exists $r'_0 > 0$ such that $P(\mathcal{A}_{\mathcal{B}}^r) = 1$ for all $r \geq r'_0$. Then, it follows that $P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}_{\mathcal{B}}^r \cap \mathcal{A}_{[n,\rho]}^r)$ is continuous when $r \geq r'_0$. In particular, $P(\mathcal{A}_{[n,\rho]}^r)$ is continuous at r'_0 . So, there is $\delta > 0$ such that $P(\mathcal{A}_{[n,\rho]}^r)$ is continuous upon $[r'_0 - \delta, r'_0 + \delta]$. Continuing this argument, continuity of $P(\mathcal{A}_{[n,\rho]}^r)$ extends until $r'_0 - \delta = 0$ so that $P(\mathcal{A}_{[n,\rho]}^r)$ is continuous for all $r \geq 0$. Q.E.D.

Theorem 11: [57, Thm. (3.8)] Suppose $\{r_k\}_{k \geq 1}$ is a sequence of radii such that $0 < r_k \leq R$ for some $R > 0$ and $\{\eta_k\}_{k \geq 1}$ is a sequence of random variables which take values r_k with probability 1. If $0 < r \leq R$ and η is a random variable taking the value r with probability 1 such that $\eta_k \Rightarrow \eta$ as $k \rightarrow \infty$. Then, $P(\mathcal{A}^{\eta_k}) \rightarrow P(\mathcal{A}^\eta)$ as $k \rightarrow \infty$.

Proof: This is just a restatement of [57, Thm. (3.8)] for the special case of random variables η_k and η such that $P(\eta_k = r_k) = 1 = P(\eta = r)$ for all $k \geq 1$. ■

Corollary 12: (to Theorem 11) $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r .

Proof: Continuity of $P(\mathcal{A}^r)$ in r follows from thm. (11). Therefore, the result follows by the discussion preceding the

statement of thm. (11). ■

Theorem 13: $r = r(n, \rho, \epsilon)$ is a continuous function of ϵ if and only if $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r .

Proof: Suppose $r(n, \rho, \epsilon)$ is a continuous function of ϵ and let $\{\epsilon_k \in (0, \frac{1}{2})\}_{k \geq 1}$ be a sequence of positive real numbers such that $\epsilon_k \rightarrow \epsilon_0$ as $k \rightarrow \infty$. Let $\{X(e)\}_{e \in G(\mathcal{X}_n, r)}$ be a finite sequence of uniformly distributed random variables with values in $[0, 1]$ and define a sequence of random variables $\{\eta_k\}_{k \geq 1}$ by $\eta_k(e) = r(n, \rho, \epsilon_k) \equiv r_k$ when $X(e) < 1$ and 0 otherwise. Clearly, $\eta_k = r_k$ with probability 1 for all $k \geq 1$. Likewise, define a random variable η_0 by $\eta_0(e) = r(n, \rho, \epsilon_0) \equiv r_0$ when $X(e) < 1$ and 0 otherwise so that $\eta_0 = r_0$ with probability 1. Since $r(n, \rho, \epsilon)$ is continuous in ϵ , then $r_k \rightarrow r_0$ as $k \rightarrow \infty$ so that $\eta_k \Rightarrow \eta_0$ as $k \rightarrow \infty$. Now, define $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$. By lemma (82), $0 < \eta_k \leq R$ for all $k \geq 0$. Therefore, $P(\mathcal{A}_{[n,\rho]}^{\eta_k}) \rightarrow P(\mathcal{A}_{[n,\rho]}^{\eta_0})$ as $k \rightarrow \infty$ by cor. (12) since $r_k \rightarrow r_0$ as $k \rightarrow \infty$. Thus, $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r . Conversely, suppose $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r and let $\{\epsilon_k \in (0, \frac{1}{2})\}_{k \geq 1}$ be any convergent sequence such that $\epsilon_k \rightarrow \epsilon_0$. Define $r_k = r(n, \rho, \epsilon_k)$ and $r_0 = r(n, \rho, \epsilon_0)$. Given $\xi > 0$, it is true that $\Xi \equiv \{k \geq 1 : |P(\mathcal{A}_{[n,\rho]}^{r_k}) - P(\mathcal{A}_{[n,\rho]}^{r_0})| \geq \xi\}$ is a set of measure zero by the continuity assumption. Therefore, $r_k \rightarrow r_0$ as $k \rightarrow \infty$ by prop. (83). Thus, suppose that $\Xi \neq \emptyset$. Then, Ξ is at most countable so that $\Xi = \emptyset$ a.s. Hence, $r_k \rightarrow r_0$ as $k \rightarrow \infty$ by prop. (83) so that $r(n, \rho, \epsilon)$ is a continuous function of ϵ . ■

G. Continuum Giant Component

Theorem 14: There exists $r_0 = r_0(n, \rho, \epsilon) < \infty$, independent of ϵ , such that

$$P(\mathcal{A}_{[n,\rho]}^{r_0}) = \frac{1}{2}.$$

Proof: Let $\epsilon \in (0, \frac{1}{2})$ be given. Since $\mathcal{A}_{[n,\rho]}^r$ is an increasing property in r by prop. (73), thm. (7) applies. Thus, there exists an interval I_ϵ of length $\Delta(n, \rho, \epsilon)$ such that $P(\mathcal{A}_{[n,\rho]}^r) \in [\epsilon, 1 - \epsilon]$ for $r \in I_\epsilon$. Since $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of r by cor. (12) and non-decreasing in r by prop. (76) and $\frac{1}{2} \in [\epsilon, 1 - \epsilon]$, then there exists $r_0 \in I_\epsilon$ such that $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1/2$. If $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$, then by lemma (82), it is true that $0 < r_0(n, \rho, \epsilon) \leq R < \infty$. It remains to be shown that $r_0 = r_0(n, \rho, \epsilon)$, independent of ϵ . ■

Recall that $\rho \in (\frac{1}{2}, 1)$ and note that the maximum distance between any two connected points in \mathcal{B} is inversely proportional to n . Then, the particular r_0 , which meets the requirements of thm. (14), is the exact radius, such that, it is equally probable that more than half of all points are connected contiguously, in which case, only one such cluster exists, with all other clusters being disjoint and sparsely connected throughout \mathcal{B} , or all connected clusters disjointly contain half (or less than half) of all available points, in which case, more than one such cluster can exist. As such, r_0 demarcates the radial connection length at which the property $\mathcal{A}_{[n,\rho]}^r$ undergoes a phase transition such that the graph $G(\mathcal{X}_n; r)$ is likely to be sparsely connected and form disjoint connected classes of points [57, Thms. (3.3, 3.6)], almost surely, when $r \in [0, r_0]$,

while $G(\mathcal{X}_n; r)$ is more likely to be fully connected and form one connected class of points [57, *Thms.* (3.3, 3.6)], almost surely, when $r \in (r_0, 1]$.

Lemma 15: $r_0 = r_0(n, \rho, \epsilon)$ is independent of ϵ .

Proof: Let $\epsilon_1, \epsilon_2 \in (0, \frac{1}{2})$ and suppose $r_{0,1} = r_0(n, \rho, \epsilon_1), r_{0,2} = r_0(n, \rho, \epsilon_2)$ such that

$$P(\mathcal{A}_{[n,\rho]}^{r_{0,1}}) = \frac{1}{2} = P(\mathcal{A}_{[n,\rho]}^{r_{0,2}}). \quad (7)$$

It has to be shown that $r_{0,1} = r_{0,2}$. Let $\{\epsilon_k\}_{k \geq 1}$ be a sequence such that $\epsilon_k = \epsilon_1$ for all $k \geq 1$ and define $r_{0,k} = r_0(n, \rho, \epsilon_k)$. Then, for arbitrary $\xi > 0$, it is true that

$$\Xi \equiv \{k \geq 1 : |P(\mathcal{A}_{[n,\rho]}^{r_{0,k}}) - P(\mathcal{A}_{[n,\rho]}^{r_{0,2}})| \geq \xi\} = \emptyset \quad (8)$$

since $r_{0,k} = r_{0,1}$ for all $k \geq 1$. Hence, by prop. (83), $r_{0,k} \rightarrow r_{0,2}$ as $k \rightarrow \infty$. But, $r_{0,k} = r_{0,1}$ for all $k \geq 1$ so that $r_{0,1} = r_{0,2}$. Thus, $r_0 = r_0(n, \rho)$, independent of ϵ . ■

Remark 16: As a result of thm. (15), $r(\epsilon)$ is independent of $\epsilon > 0$, since $r(n, \rho, \epsilon) \rightarrow r(\epsilon)$ as $E[n] \rightarrow \infty$. As such, $\Delta(\epsilon) = o(1)$ so that \mathcal{A}^r has a sharp threshold, by definition (6).

Corollary 17: The critical radius, associated with the property \mathcal{A}^r , is unique.

Proof: $r(\epsilon)$ is the limit of $r(n, \rho, \epsilon)$ as $E[n] \rightarrow \infty$. As such, r_0 is the constant limit of $r_0(n, \rho)$ as $E[n] \rightarrow \infty$. ■

Corollary 18: Given $r > 0$, there exists a density of points $\lambda_0 = \lambda(n_0)$ such that

$$P(\mathcal{A}_{[n_0,\rho]}^r) = \frac{1}{2}.$$

Proof: By lemma (15), let $n_0 = n_0(r, \rho)$ be the minimum of all positive (real) solutions to $r = r_0(n, \rho)$ for some fixed $r > 0$. The result follows. ■

Since n is inversely proportional to connection distance r (requiring that $n \in [1, \infty)$), then the particular n_0 , which meets the requirements of cor. (18), is the exact number of points, such that, it is equally probable that more than half of all points are connected contiguously. In this case, only one such cluster exists, with all other clusters being disjoint and sparsely connected throughout \mathcal{B} . Otherwise, all connected clusters disjointly contain half (or less than half) of all available points, in which case, more than one such cluster can exist. As such, n_0 demarcates the number of points at which the property $\mathcal{A}_{[n,\rho]}^r$ undergoes a phase transition such that the graph $G(\mathcal{X}_n; r)$ is likely to be sparsely connected and form disjoint connected classes of points [57, *Thms.* (3.3, 3.6)], almost surely, when $n \in [1, n_0]$. Alternatively, $G(\mathcal{X}_n; r)$ is more likely to be fully connected and form one connected class of points [57, *Thms.* (3.3, 3.6)], almost surely, when $n \in (n_0, \infty)$.

H. Continuum Sharp Threshold Interval Length

Given the particular radius guaranteed by thm. (14), then thm. (7) can be used to find an estimate of the length of the sharp threshold interval such that $P(\mathcal{A}_{[n,\rho]}^r)$ increases sharply from some $\epsilon \in (0, \frac{1}{2})$ to $1 - \epsilon$. By lemma (15), it is true that

r_0 is independent of any particular ϵ . Thus, the interval and its length must be fixed, given n and $\rho \in (\frac{1}{2}, 1)$.

Theorem 19: $\Delta(n, \rho) = \Theta(r_0 \log^{\frac{1}{4}} n)$.

Proof: For $\delta \in (0, \frac{1}{2})$, let $\epsilon_\delta = \frac{1}{2} - \delta$. By thm. (7) and thms. (14) and (15),

$$\begin{aligned} \Delta(n, \rho) &= \lim_{\delta \rightarrow 0^+} \Delta(n, \rho, \epsilon_\delta) \\ &= \lim_{\delta \rightarrow 0^+} \Theta(r(n, \rho, \epsilon_\delta) \log^{\frac{1}{4}} n) \\ &= \Theta(r_0 \log^{\frac{1}{4}} n). \end{aligned}$$

■

Theorem (19) gives an expected result, given thm. (7) above. However, in [13], a much more practical estimate of this length is obtained after the bounded region is partitioned by hexagons of a known size. If M^2 is the number of these hexagons in the bounded region, then it is shown that a good estimate of the sharp threshold interval length is a polynomial in $1/M$.

Theorem 20: There is a constant $c > 0$, independent of M , such that for all $\epsilon_1 > 0$ and every fixed small $\delta > 0$

$$P(\mathcal{A}_{[n,\rho+\delta]}^r) \leq (\frac{1}{2} + \epsilon_1) M^{-c(r_0-r)} \quad (9)$$

for all $r \leq r_0$ and

$$P(\mathcal{A}_{[n,\rho-\delta]}^r) \geq 1 - (\frac{1}{2} + \epsilon_1) M^{-c(r-r_0)} \quad (10)$$

for all $r \geq r_0$.

Theorem 21: $P(\mathcal{A}_{[n,\rho]}^r)$ is a continuous function of ρ .

Remark 22: The proof of thm. (21) requires thm. (20) which will be proven later. For now, the result of thm. (21) is assumed. By thm. (21), for small $\delta > 0$,

$$P(\mathcal{A}_{[n,\rho-\delta]}^r) \approx P(\mathcal{A}_{[n,\rho]}^r) \approx P(\mathcal{A}_{[n,\rho+\delta]}^r).$$

Theorem (20) asserts that if $r_1 < r_0 < r_2$ and for some $\epsilon \in (0, \frac{1}{2})$, it is true that $P(\mathcal{A}_{[n,\rho]}^{r_1}) = \epsilon$ and $P(\mathcal{A}_{[n,\rho]}^{r_2}) = 1 - \epsilon$, then $r_2 - r_1$ is an estimate of the sharp threshold interval length for the property, $\mathcal{A}_{[n,\rho]}^r$. Later, a similar result will be stated and proven which can be used in the proof of thm. (21).

III. HEXAGONAL PARTITION MODEL

It was seen in section (II-G) that $r_0 > 0$ exists such that the probability is $1/2$ for the occurrence of the property that at least half of all points connect in the bounded region, \mathcal{B} . By thm. (7),

$$r_c = r_c(n) = O\left(\sqrt{\frac{\log n}{n}}\right) \leq r_0(n) = r_0 \quad (11)$$

where r_c defines the critical radius at which the continuum property occurs with arbitrarily small, positive probability.

For fixed $r \in (r_c, r_0)$, let h^r be the largest hexagon that can be inscribed into a circle of radius $r/4$. Let H_r be a countably infinite collection of copies of h^r such that

$$\mathbb{R}^2 = \bigcup_{h_{i,j}^r \in H_r} h_{i,j}^r \quad (12)$$

and for $h_{i,j}^r, h_{i',j'}^r \in H_r$, it is true that $h_{i,j}^r \neq h_{i',j'}^r$ whenever $|i - i'| + |j - j'| \neq 0$. Connectivity between $x, y \in \mathcal{X}_n$ is then defined as x and y both lying in the same hexagon or neighboring hexagons.

With the bounded region \mathcal{B} partitioned into hexagons contained within $\mathcal{B} \cap H_r$, the analysis proceeds whereby the original problem of estimating the sharp threshold interval length in the continuum is now replaced by the problem of estimating the length in the hexagonal partition framework. As such, definitions of connectivity and the increasing property are defined in the new framework. Then, the continuity and existence results are shown to still hold in the new framework. Later, an analogue to thm. (20) is stated and proven.

A. Definitions

Definition 23: A hexagonal partition of \mathcal{B} is a finite collection of hexagons from H_r such that \mathcal{B} is a union of all hexagons in the finite collection.

Definition 24: The Hamming distance between elements, $h_{i,j}^r, h_{i',j'}^r \in H_r$ is defined to be the quantity

$$h(h_{i,j}^r, h_{i',j'}^r) = |i - i'| + |j - j'|.$$

Definition 25: Points $x, y \in \mathcal{X}_n$ are H_r -connected and $\langle x, y \rangle_{H_r}$ is an H_r -open edge, if there exists $h_{i_x,j_x}^r, h_{i_y,j_y}^r \in H_r$ such that $x \in h_{i_x,j_x}^r$ and $y \in h_{i_y,j_y}^r$ where $h(h_{i_x,j_x}^r, h_{i_y,j_y}^r) \leq 2$ with $|i_x - i_y| \leq 1$ and $|j_x - j_y| \leq 1$. Points in \mathcal{X}_n are H_r -disconnected and form an H_r -closed edge otherwise.

Definition 26: Given a $y \in \mathcal{X}_n$, an H_r -connected component containing y is the subset of points $\langle C_y \rangle_{H_r} \subseteq \mathcal{X}_n$ containing y and every $x \in \mathcal{X}_n \setminus \{y\}$ having an H_r -open set of edges connecting x to y .

Definition 27: Given an H_r -connected edge, $e = \langle x, y \rangle_{H_r}$, an H_r -connected component containing e is the subset of points $\langle C_e \rangle_{H_r} \subseteq \mathcal{X}_n$ containing x and y and every $z \in \mathcal{X}_n \setminus \{x, y\}$ having an H_r -open set of edges connecting z to both x and y .

B. The Increasing Property

1) Bounded Number of Nodes: Let $\langle C \rangle_{H_r} \subseteq \mathcal{X}_n$ be an H_r -connected component such that $|\langle C \rangle_{H_r}| = N$ and let $\rho_n(C) = \frac{N}{n}$ be defined as in section (II-E1). For $\rho \in (\frac{1}{2}, 1)$, define the graph property of all connected components containing at least 100% of all available points by

$$\mathcal{A}_{[n,\rho]}^{H_r} = \{\langle C \rangle_{H_r} \subseteq \mathcal{X}_n : \rho_n(C) \geq \rho\}. \quad (13)$$

As in [31], for $\epsilon \in (0, \frac{1}{2})$, define

$$r^*(n, \rho, \epsilon) = \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^{H_r}) \geq \epsilon\} \quad (14)$$

to be the critical radius at which $\mathcal{A}_{[n,\rho]}^{H_r}$ occurs with probability at least ϵ and define

$$\Delta^*(n, \rho, \epsilon) = r^*(n, \rho, 1 - \epsilon) - r^*(n, \rho, \epsilon) \quad (15)$$

to be the length of the continuum of radii upon which $\mathcal{A}_{[n,\rho]}^{H_r}$ increases in probability of occurrence from $\epsilon > 0$ to $1 - \epsilon > 0$.

2) Unbounded Number of Nodes: In the event that n is unbounded, define the corresponding graph property to be

$$\mathcal{A}^{H_r} = \{\langle C \rangle_{H_r} \subseteq \mathcal{X}_\infty : |\langle C \rangle_{H_r}| = \infty\}. \quad (16)$$

Define

$$r^*(\epsilon) = \inf\{r > 0 : P(\mathcal{A}^{H_r}) \geq \epsilon\} \quad (17)$$

to be the critical radius at which \mathcal{A}^{H_r} occurs with probability at least ϵ and define

$$\Delta^*(\epsilon) = r^*(1 - \epsilon) - r^*(\epsilon) \quad (18)$$

to be the length of the continuum of radii upon which \mathcal{A}^{H_r} increases in probability of occurrence from $\epsilon > 0$ to $1 - \epsilon > 0$.

C. Continuity Results and Some Continuum Relationships

The continuity results of section (II-F) hold for the properties defined after the bounded region \mathcal{B} is partitioned by copies of the hexagon h^r , since connectivity is now characterized by points lying within distance $r/2$ (within neighboring hexagons). As such, the hexagonal partition connectivity model is only another way of viewing the continuum connectivity model. Then, by thm. (14), there exists $r_0^* = r_0^*(n, \rho)$ which satisfies the criteria of the theorem for the property $\mathcal{A}_{[n,\rho]}^{H_r}$.

Definition 28: $G(\mathcal{X}_n; H_r)$ is defined to be the H_r -graph of all H_r -open and H_r -closed edges between points in $\mathcal{X}_n \subset \mathcal{B}$.

In addition to the continuity results under r -connectivity also holding under H_r -connectivity, the next lemma shows that the graph of the set of clusters formed under H_r -connectivity is a sub-graph of the set of clusters formed under r -connectivity.

Lemma 29: $G(\mathcal{X}_n; H_r) \subseteq G(\mathcal{X}_n; r)$.

Proof: Suppose $\langle x, y \rangle_{H_r} \in G(\mathcal{X}_n; H_r)$ is any H_r -connected edge. Without loss of generality, choose a coordinate system on \mathbb{R}^2 so that $\langle x, y \rangle_{H_r}$ lies on a coordinate axis with $\hat{0} = (0, 0)$ defined such that $d(x, \hat{0}) = \frac{d(x,y)}{2} = d(\hat{0}, y)$. Since $x, y \in \mathcal{X}_n \subset \mathcal{B}$ and H_r is a partition of \mathcal{B} , then there exists $h_{i_x,j_x}^r, h_{i_y,j_y}^r \in H_r$ such that $x \in h_{i_x,j_x}^r, y \in h_{i_y,j_y}^r$ and $h(h_{i_x,j_x}^r, h_{i_y,j_y}^r) \leq \max\{|i_x - i_y|, |j_x - j_y|\} \leq 1$. Each of h_{i_x,j_x}^r and h_{i_y,j_y}^r are copies of h^r and can be inscribed into copies of a circle of radius $\frac{r}{4}$. Therefore, $d(x, y) = d(x, \partial h_{i_x,j_x}^r) + d(\partial h_{i_y,j_y}^r, y) \leq \frac{r}{2} + \frac{r}{2} = r$ so that $x, y \in \mathcal{X}_n$ are r -connected. Thus, $\langle x, y \rangle_{H_r} \in G(\mathcal{X}_n; r)$, which shows that $G(\mathcal{X}_n; H_r) \subseteq G(\mathcal{X}_n; r)$. ■

Using lemma (29), the next results show that given a sample of size $n > 1$ and a connectivity radius $r > 0$, the probability of the event of one subset of connected data points containing 100% of the n data points, for $\rho \in (\frac{1}{2}, 1)$ is (possibly) smaller under H_r -connectivity than under r -connectivity. In addition, a (possibly) larger radius of connectivity is required to achieve the same proportion of data points being connected into one cluster.

Lemma 30: $P(\mathcal{A}_{[n,\rho]}^{H_r}) \leq P(\mathcal{A}_{[n,\rho]}^r)$.

Proof: By lemma (29), it is true that $\mathcal{A}_{[n,\rho]}^{H_r} \subseteq \mathcal{A}_{[n,\rho]}^r$. ■

Lemma 31: $r_0 \leq r_0^*$.

Proof: Seeking a contradiction, suppose $r_0 > r_0^*$. Then,

$$\frac{1}{2} = P(\mathcal{A}_{[n,\rho]}^{r_0}) \quad (19)$$

$$\geq P(\mathcal{A}_{[n,\rho]}^{r_0^*}) \quad (20)$$

$$\geq P(\mathcal{A}_{[n,\rho]}^{H_{r_0^*}}) \quad (21)$$

$$= \frac{1}{2} \quad (22)$$

where equality (19) follows by thm. (14), ineq. (20) follows by properties of probability measures and by hypothesis, ineq. (21) follows by lemma (30) and equality (22) follows by thm. (14). It follows that $P(\mathcal{A}_{[n,\rho]}^{r_0^*}) = 1/2$. Therefore, $r_0^* \in \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) = \frac{1}{2}\}$ and $r_0^* < r_0 = \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) = \frac{1}{2}\}$. This is a contradiction. Thus, $r_0 \leq r_0^*$. ■

Later, it will be shown that equality holds for both lemmas (30,31) under special conditions which are perfect for K -means clustering.

D. Hexagonal Sharp Threshold Interval Length

Given the particular radius guaranteed by thm. (14), then thm. (7) can be used to find an estimate of the length of the sharp threshold interval such that $P(\mathcal{A}_{[n,\rho]}^{H_r})$ increases sharply from some $\epsilon \in (0, \frac{1}{2})$ to $1 - \epsilon$. By lemma (15), it is true that r_0^* is independent of any particular ϵ . Thus, the interval and its length must be fixed given n and $\rho \in (\frac{1}{2}, 1)$.

Theorem 32: $\Delta^*(n, \rho) = \Theta(r_0^* \log^{\frac{1}{4}} n)$.

Proof: For $\delta \in (0, \frac{1}{2})$, let $\epsilon_\delta = \frac{1}{2} - \delta$. By thm. (7) and thms. (14) and (15),

$$\begin{aligned} \Delta^*(n, \rho) &= \lim_{\delta \rightarrow 0^+} \Delta^*(n, \rho, \epsilon_\delta) \\ &= \lim_{\delta \rightarrow 0^+} \Theta(r^*(n, \rho, \epsilon_\delta) \log^{\frac{1}{4}} n) \\ &= \Theta(r_0^* \log^{\frac{1}{4}} n). \end{aligned}$$

As in thm. (19) above, thm. (32) gives an expected result, given thm. (7) above. Likewise, a similar result to [13, Thm. (3.3.1)] can be stated and later proven, as in the case of thm. (20). It is the result of thm. (33) that allows us to estimate the length of the sharp threshold interval in the presence of the hexagonal partition of \mathcal{B} .

Theorem 33: There is a constant $c > 0$, independent of M , such that for all $\epsilon_1 > 0$ and every fixed small $\delta > 0$

$$P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) \leq \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r_0^*-r)}$$

for all $r \leq r_0^*$ and

$$P(\mathcal{A}_{[n,\rho-\delta]}^{H_r}) \geq 1 - \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r-r_0^*)} \quad (23)$$

for all $r \geq r_0^*$.

Let M^2 be the number of hexagons partitioning the region \mathcal{B} and let $H_{\mathcal{B}}(r) = H_r \cap \mathcal{B}$. Given $\langle C \rangle_{H_r} \subseteq \mathcal{X}_n$, define $H_C = \{h_{\mathcal{B}}^r \in H_{\mathcal{B}}(r) : h_{\mathcal{B}}^r \cap \langle C \rangle_{H_r} \neq \emptyset\}$ to be

the connected cluster of hexagons such that each hexagon contains at least one point from the connected cluster of points, $\langle C \rangle_{H_r}$.

Lemma 34: $E[\rho_n(C)] = \frac{E[|H_C|]}{M^2}$.

Proof: Let $\langle C \rangle_{H_r} \subseteq \mathcal{X}_n$ be an H_r -connected cluster and let K_{H_C} be a random variable taking as values the number of points in the region R_{H_C} defined by the hexagons in H_C . Since the n points are uniformly distributed spatially and \mathcal{B} is partitioned into M^2 copies of the prototypical hexagon h^r , then

$$\begin{aligned} E[K_{H_C}] &= n \frac{E[\text{area}(R_{H_C})]}{\text{area}(\mathcal{B})} \\ &= n \frac{E[|H_C|] \times \text{area}(h^r)}{M^2 \times \text{area}(h^r)} \\ &= n \frac{E[|H_C|]}{M^2}. \end{aligned}$$

But, $E[K_{H_C}] = E[\langle C \rangle_{H_r}]$. Therefore,

$$E[\langle C \rangle_{H_r}] = n \frac{E[|H_C|]}{M^2}$$

implies

$$E[\rho_n(C)] = \frac{E[|H_C|]}{M^2}.$$

■

Define $\mathcal{D}_{[n,\rho]}^r = \{H_C \subseteq H_{\mathcal{B}}(r) : E[\rho_n(C)] \geq \rho\}$. With $\mathcal{D}_{[n,\rho]}^r$ defined as such, the original problem of estimating the length of the sharp threshold for the property $\mathcal{A}_{[n,\rho]}^r$ in the continuum is now recast as a site percolation problem on a hexagonal lattice. As will be defined later, a site in the lattice will be deemed open if the corresponding hexagon is occupied by at least one of the points from \mathcal{X}_n and it will be deemed closed otherwise. Likewise, two sites are connected and belong to the same connected cluster if both sites are open and their hamming distance is less than or equal to one. Later, a torus on the lattice will be formed by defining a countable collection of permutations of the hexagons in the partition so that the length of the sharp threshold for the property $\mathcal{D}_{[n,\rho]}^r$ can be approximated by the length for another property $\tilde{\mathcal{D}}_{[n,\rho]}^r$ on the torus. In this way, boundary connection issues for sites in the partition of \mathcal{B} are mitigated and the length of the sharp threshold interval for the property $\tilde{\mathcal{D}}_{[n,\rho]}^r$ approximates the length for $\mathcal{D}_{[n,\rho]}^r$, which approximates the length for $\mathcal{A}_{[n,\rho]}^{H_r}$, which finally approximates the length for $\mathcal{A}_{[n,\rho]}^r$, the original property in the continuum.

Theorem 35: There is a constant $c > 0$, independent of M , such that

$$P(\mathcal{D}_{[n,\rho]}^r) \leq \frac{1}{2} M^{-c(r_0^*-r)}$$

for all $r \leq r_0^*$. Similarly, for some fixed small $\delta > 0$ and for all $\epsilon_1 > 0$, there is an $M_0(\delta, \epsilon_1)$ such that for all $M > M_0(\delta, \epsilon_1)$

$$P(\mathcal{D}_{[n,\rho-\delta]}^r) \geq 1 - \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r-r_0^*)}$$

for all $r \geq r_0^*$.

An important part of the proof of thm. (35) relies upon the sharp threshold inequality results of [12] and [26]. In order to apply these results, connectivity in the hexagon lattice structure should be extended to the case of a torus, whereby any boundary connectivity issues are mitigated. As such, make $H_{\mathcal{B}}(r)$ into a torus by identifying $h_{i,j} \in H_{\mathcal{B}}(r)$ with an element $h_{i',j'}$ in a copy of $H_{\mathcal{B}}(r)$, if $i' = i \bmod M$ and $j' = j \bmod M$. For every $k, l \in \mathbb{Z}$, the mapping $\tau_{k,l} : h_{i,j} \rightarrow h_{i+k,j+l}$ defines a shift translation. In this way, a subgroup of automorphisms $\tau = \{\tau_{k,l} : k, l \in \mathbb{Z}\}$ with the transitivity property is formed. Thus, any hexagon $h_{i,j}$ can be shifted to any other hexagon $h_{i',j'}$ with the translation, $\tau_{i'-i,j'-j}$. Now, hexagons in the 1st row (column) are allowed to be joined in a connected cluster with hexagons in the Mth row (column), provided that all hexagons in question are occupied.

Proposition 36: Define $\tau(H_{\mathcal{B}}(r))$ to be the torus created by translations of hexagons in $H_{\mathcal{B}}(r)$ under the action of permutations in τ and define $\hat{\mathcal{D}}_{[n,\rho]}^r = \{H_C \subseteq \tau(H_{\mathcal{B}}(r)) : E[\rho_n(C)] \geq \rho\}$. Then, $\mathcal{D}_{[n,\rho]}^r \subset \hat{\mathcal{D}}_{[n,\rho]}^r$ and $\mathcal{D}_{[n,\rho]}^r \neq \hat{\mathcal{D}}_{[n,\rho]}^r$.

Proof: Since $\hat{\mathcal{D}}_{[n,\rho]}^r$ contains all of the connected hexagons from $\mathcal{D}_{[n,\rho]}^r$ and any connections between the 1st and Mth rows (columns) while $\mathcal{D}_{[n,\rho]}^r$ contains no connection between the 1st and Mth rows (columns), then the result follows. ■

Definition 37: To each hexagon in the partition of \mathcal{B} , associate a site $i \in \{1, 2, \dots, M^2\}$ as the center of the hexagon. For sites $i \in \{1, 2, \dots, M^2\}$, define $s_i \in \{0, 1\}$ to be the state on site i . A site i is said to be *open* if $s_i = 1$ and *closed* otherwise. There exists an edge $e_{\{i,j\}}$ between sites $i, j \in \{1, 2, \dots, M^2\}$ if and only if there exists a hexagon $h_{i,j}^r \ni i, j$ or there exists neighboring hexagons $h_i^r \ni i$ and $h_j^r \ni j$ in the partition of \mathcal{B} . Define $e_{\{i,j\}}$ to be *open* if and only if $s_i = 1 = s_j$ and *closed* otherwise.

Definition 38: The conditional influence of i on the property $\hat{\mathcal{D}}_{[n,\rho]}^r$ is defined to be

$$I(i) = P(\hat{\mathcal{D}}_{[n,\rho]}^r \mid s_i = 1) - P(\hat{\mathcal{D}}_{[n,\rho]}^r \mid s_i = 0)$$

and it is a measure of the change in the probability of $\hat{\mathcal{D}}_{[n,\rho]}^r$ due to a state change from $s_i = 0$ to $s_i = 1$ at site, i .

For completeness, [13, Lemma (4.1.1)] is stated without proof, which gives an upper bound on the change in $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ as a function of the point density λ . Utilizing the chain rule for derivatives, a lower bound on the change in $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ as a function of r is found and the resulting inequality relationship is used to estimate upper and lower bounds on $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$, which will approximate the inequality results of thm. (35).

Lemma 39: [13, Lemma (4.1.1)] There is a constant $z > 0$, independent of M and λ , such that

$$\frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq z^*(\lambda) \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \log M$$

where A_{h^r} is the area of the prototypical hexagon h^r and $z^*(\lambda) = -zA_{h^r}e^{-A_{h^r}\lambda}$.

Lemma 40: There is a constant $c > 0$, independent of M and λ , such that

$$\frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq c^*(\lambda) \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \log M$$

where A_{h^r} is the area of the prototypical hexagon h^r and $c^*(\lambda) = c(\lambda)A_{h^r}e^{-A_{h^r}\lambda}$, with $c(\lambda) = -cg(\lambda)$ for some function $g(\lambda)$.

Proof: As in cor. (18), let n^* be the inverse of r^* and seeking a contradiction, suppose $dr/d\lambda = 0$. Let $\epsilon \in (0, \frac{1}{2})$. By lemma (39), $dP/d\lambda$ exists. Now, the existence of dP/dr will be shown by proving a Lipschitz condition on the probability distribution $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ as a function of r . Assume $\text{area}(\mathcal{B}) = 1$. Without loss of generality, it can be assumed that $r \in [0, 1]$. Without further loss of generality, let $r_1^*, r_2^* \in [0, 1]$ such that r_0^* is the midpoint of $[r_1^*, r_2^*]$, i.e. $r_0^* = (r_2^* - r_1^*)/2$. Then, by thm. (32),

$$|P(\hat{\mathcal{D}}_{[n,\rho]}^{r_2^*}) - P(\hat{\mathcal{D}}_{[n,\rho]}^{r_1^*})| \leq 1 = (\Delta^*(n, \rho))^{-1} |r_2^* - r_1^*|.$$

Therefore, $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is Lipschitz continuous with respect to r . Hence, dP/dr exists. Now, since $dP/d\lambda$, dP/dr and $dr/d\lambda$ all exist, then the Chain Rule for derivatives yields,

$$\frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) = \frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \times \frac{dr}{d\lambda}.$$

Note that the existence of dP/dr requires that $|dP/dr| < \infty$. Therefore, since $dr/d\lambda = 0$, then

$$\frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) = \frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \times 0 = 0.$$

As a result, $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is constant as a function of λ . So, suppose that $0 < n < n^*$. Then, $P(\hat{\mathcal{D}}_{[n,\rho]}^r) = 0$, which implies that $P(\hat{\mathcal{D}}_{[n,\rho]}^r) \equiv 0$. This is a contradiction, since $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is a probability distribution. Hence, $dr/d\lambda \neq 0$. Now, by [33, Thm. (2.28)], there is a constant $c > 0$, independent of M and λ , such that

$$I(i) \geq c \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \frac{\log M}{M^2}.$$

Under the action of τ , each hexagon in the bounded region \mathcal{B} is translated to another hexagon in a copy of \mathcal{B} . Therefore, $\hat{\mathcal{D}}_{[n,\rho]}^r$ and $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ are invariant under the action of τ . Hence, $I(i) = I(j)$ whenever, $\tau(i) = j$, where $\tau(i)$ is defined to be the translation of the hexagon $h_i^r \ni i$ to the hexagon $h_j^r \ni j$ in the copy of the partition of \mathcal{B} . From [13], in the proof of thm. (39), the following identity holds, with $p = p(\lambda) = 1 - e^{-A_{h^r}\lambda}$ defined above,

$$\begin{aligned} \frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) &= \frac{d}{dp} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \times \frac{dp}{d\lambda} \\ &= -A_{h^r} e^{-A_{h^r}\lambda} \sum_{i=1}^{M^2} I(i). \end{aligned} \quad (24)$$

For $\gamma > 0$, $r > 0$ and $k > 0$, any H_r -connected component in \mathcal{X}_n containing at least $\gamma(n+k)/2$ points will inherently contain an H_r -connected component of size at least $\gamma n/2$. Hence, $\mathcal{A}_{[\gamma(n+k), \rho]}^{H_r} \subseteq \mathcal{A}_{[\gamma n, \rho]}^{H_r}$. It follows that $P(\mathcal{A}_{[\gamma(n+k), \rho]}^{H_r}) \leq P(\mathcal{A}_{[\gamma n, \rho]}^{H_r})$. Therefore, $r^*(\gamma n, \rho, \epsilon) \in \{r > 0 : P(\mathcal{A}_{[\gamma(n+k), \rho]}^{H_r}) \geq \epsilon\}$, which implies $r^*(\gamma(n+k), \rho, \epsilon) \leq r^*(\gamma n, \rho, \epsilon)$ for $k > 0$. Hence,

$$r^*(\gamma(n+k), \rho, \epsilon) - r^*(\gamma n, \rho, \epsilon) \leq 0. \quad (25)$$

Since point density λ is proportional to point count n for any bounded region \mathcal{B} , then using ineq. (25) yields

$$\frac{dr}{d\lambda} = \lim_{k \rightarrow 0} \frac{r^*(\gamma(n+k), \rho, \epsilon) - r^*(\gamma n, \rho, \epsilon)}{\gamma k} \leq 0,$$

for some $\gamma > 0$. Since $dr/d\lambda \neq 0$, it follows that

$$\frac{dr}{d\lambda} < 0.$$

Since $dr/d\lambda$ exists, then $|dr/d\lambda| < \infty$. Thus, by substituting

$$I(i) \geq c \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \frac{\log M}{M^2}$$

into (24), it follows that

$$\begin{aligned} \frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) &= -A_{hr} e^{-A_{hr}\lambda} \sum_{i=1}^{M^2} I(i) \\ &\leq -c A_{hr} e^{-A_{hr}\lambda} \\ &\times \sum_{i=1}^{M^2} \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \frac{\log M}{M^2} \\ &= -c A_{hr} e^{-A_{hr}\lambda} \\ &\times \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \log M. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{d}{d\lambda} P(\hat{\mathcal{D}}_{[n,\rho]}^r) &= \frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \times \frac{dr}{d\lambda} \\ &\leq -c A_{hr} e^{-A_{hr}\lambda} \end{aligned} \quad (26)$$

$$\times \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \quad (27)$$

$$\times \log M \quad (28)$$

so that

$$\frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq -c A_{hr} e^{-A_{hr}\lambda} \left(\frac{dr}{d\lambda} \right)^{-1} \quad (29)$$

$$\times \min\{P(\hat{\mathcal{D}}_{[n,\rho]}^r), 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)\} \quad (30)$$

$$\times \log M. \quad (31)$$

Defining $g(\lambda) = (dr/d\lambda)^{-1}$, the result follows. ■

Remark 41: Let $\epsilon > 0$ be given. At the risk of ambiguity, denote $n = E[n]$ and define $\lambda^*(n, \rho, \epsilon) = \inf \{n > 0 \mid P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq \epsilon\}$. Inequality (26) implies that $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is increasing as a function of decreasing node density $\lambda = \lambda^*(n, \rho, \epsilon)$ such that the event $\{P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq \epsilon\}$ first occurs. Likewise, since the maximum distance between connected points is inversely proportional to node density, then ineq. (29) implies that $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is decreasing as a function of increasing maximum distance $r = r^*(n, \rho, \epsilon)$ between connected points such that the event $\{P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq \epsilon\}$ first occurs.

Lemma 42: Let $c > 0$ be as in thm. (40). Then, there exists r_0^* , independent of M , such that

$$P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq \frac{1}{2} M^{-c(r_0^* - r)}$$

for all $r \leq r_0^*$ and

$$P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq 1 - \frac{1}{2} M^{-c(r - r_0^*)}$$

for all $r \geq r_0^*$.

Proof: Arguing as in the proof to thm. (14), there exists r_0^* such that $P(\hat{\mathcal{D}}_{[n,\rho]}^{r_0^*}) = 1/2$. Arguing similarly to cor. (12), $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ is continuous in r . Therefore, $P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ for $r \leq r_0^*$ and $P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq 1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ for $r \geq r_0^*$. Thus, the result of lemma (40) takes the form

$$\frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq c^*(\lambda) P(\hat{\mathcal{D}}_{[n,\rho]}^r) \log M$$

for $r \leq r_0^*$ and

$$\frac{d}{dr} P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq c^*(\lambda) (1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)) \log M$$

for $r \geq r_0^*$. The last two inequalities can be written

$$\frac{d}{dr} \log P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq c^*(\lambda) \log M$$

for $r \leq r_0^*$ and

$$\frac{d}{dr} \log (1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)) \leq -c^*(\lambda) \log M$$

for $r \geq r_0^*$, respectively. Consider $r \leq r_0^*$. Both sides of

$$\frac{d}{dr} \log P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq c^*(\lambda) \log M$$

are integrated in the direction of increasing point density since $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ decreases as a function of point density λ by the proof to lemma (40). It was also shown that $dr/d\lambda < 0$, i.e. r is decreasing as a function of point density. Therefore, the integration limits for the interval $[r, r_0^*]$ are from r_0^* to r . Noting that the inequality is reversed for backward integration, the following is obtained for $c > 0$ and some $K_1(\lambda) \geq 0$,

$$\log P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq K_1(\lambda) \log M^{c(r_0^* - r)}$$

which can be rewritten as

$$\log P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq K_1(\lambda) \log M^{-c(r_0^* - r)}.$$

This implies

$$P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq K_2(\lambda) M^{-c(r_0^* - r)}$$

for some $K_2(\lambda) \geq 0$. Therefore, using the initial condition $P(\hat{\mathcal{D}}_{[n,\rho]}^{r_0^*}) = 1/2$ yields $K_2(\lambda) = 1/2$. Thus,

$$P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq \frac{1}{2} M^{-c(r_0^* - r)}.$$

Now, consider $r \geq r_0^*$. Similarly, both sides of

$$\frac{d}{dr} \log (1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)) \leq -c^*(\lambda) \log M$$

are integrated in the direction of increasing connection radii on $[r_0^*, r]$ since $P(\hat{\mathcal{D}}_{[n,\rho]}^r)$ increases as a function of connection radii r by the proof to lemma (40). Therefore, the integration limits are from r_0^* to r . The following is obtained for $c > 0$ and some $K_3(\lambda) \geq 0$,

$$\log (1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)) \leq -K_3(\lambda) \log M^{c(r - r_0^*)}$$

which can be rewritten as

$$\begin{aligned} \log(1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r)) &\leq -K_3(\lambda) \log M^{-c(r_0^* - r)} \\ &= K_3(\lambda) \log M^{-c(r - r_0^*)}. \end{aligned}$$

This implies

$$1 - P(\hat{\mathcal{D}}_{[n,\rho]}^r) \leq K_4(\lambda) M^{-c(r - r_0^*)}$$

for some $K_4(\lambda) \geq 0$. Therefore, using the initial condition $P(\hat{\mathcal{D}}_{[n,\rho]}^{r_0^*}) = 1/2$ yields $K_4(\lambda) = 1/2$. Hence,

$$P(\hat{\mathcal{D}}_{[n,\rho]}^r) \geq 1 - \frac{1}{2} M^{-c(r - r_0^*)}.$$

■

By prop. (36), there are cases when $\mathcal{D}_{[n,\rho]}^r \subset \hat{\mathcal{D}}_{[n,\rho]}^r$, but $\mathcal{D}_{[n,\rho]}^r \neq \hat{\mathcal{D}}_{[n,\rho]}^r$ so that the occurrence of $\hat{\mathcal{D}}_{[n,\rho]}^r$ does not imply the occurrence of $\mathcal{D}_{[n,\rho]}^r$. To exclude these possibilities, the arguments of [13] are followed whereby a slightly larger property $\mathcal{D}_{[n,\rho-\delta]}^r$ is considered for some small $\delta > 0$ such that the occurrence of $\hat{\mathcal{D}}_{[n,\rho]}^r$ implies the occurrence of $\mathcal{D}_{[n,\rho-\delta]}^r$.

As in [13], let $\phi(M)$ be any M -dependent integer such that $\phi(M) \rightarrow \infty$ as $M \rightarrow \infty$ and

$$\phi(M) = o(c(r - r_0^*) \log M).$$

Choose a coordinate system so that \mathcal{B} has its lower left corner at the origin. Define the top, bottom, left and right boundary strips of \mathcal{B} as $H_i, i = 1, 2, 3, 4$ with sizes $\phi(M) \times M, \phi(M) \times M, M \times \phi(M)$ and $M \times \phi(M)$ by

$$H_1 = \{H_{i,j} : i = M - \phi(M) + 1, \dots, M, j = 1, \dots, M\}$$

$$H_2 = \{H_{i,j} : i = 1, \dots, \phi(M), j = 1, \dots, M\}$$

$$H_3 = \{H_{i,j} : i = 1, \dots, M, j = 1, \dots, \phi(M)\}$$

$$H_4 = \{H_{i,j} : i = 1, \dots, M, j = M - \phi(M) + 1, \dots, M\}.$$

Let E_i be the event that there is a connected path of *occupied* hexagons crossing the rectangle H_i using the longest straight-line path.

Lemma 43: For $i = 1, 2, 3, 4$, there are constants $c_i > 0$ such that for large M and $r \geq r_0^*$

$$P(E_i) \geq 1 - e^{-c_i \phi(M)}.$$

Proof: As in [13], by the duality property, the occurrence of $E_i, i = 1, 2, 3, 4$ is equivalent to the non-occurrence of the event that there is a connected path of *unoccupied* hexagons crossing $H_i, i = 1, 2, 3, 4$ using the shortest straight-line path. The rest of the proof follows [13] with the edge probability as a function of point density $p(\lambda_0)$ replaced by $r^*(n, \rho, \epsilon)$ and the critical probability for the occurrence of an infinite cluster of occupied hexagons p_c replaced by r_0^* . ■

Proof: (Theorem 35) By prop. (36), $\mathcal{D}_{[n,\rho]}^r \subset \hat{\mathcal{D}}_{[n,\rho]}^r$ so that $P(\mathcal{D}_{[n,\rho]}^r) \leq P(\hat{\mathcal{D}}_{[n,\rho]}^r)$. To estimate $P(\mathcal{D}_{[n,\rho-\delta]}^r)$ for $r > r_0$ and any given $\delta > 0$, let $E = E_1 \cap E_2 \cap E_3 \cap E_4$ and

consider $F = \hat{\mathcal{D}}_{[n,\rho]}^r \cap E$. Since $P(F) = P(F \cap \mathcal{D}_{[n,\rho-\delta]}^r) + P(F - \mathcal{D}_{[n,\rho-\delta]}^r)$, then

$$P(\mathcal{D}_{[n,\rho-\delta]}^r) \geq P(F) - P(F - \mathcal{D}_{[n,\rho-\delta]}^r).$$

Noting that $P(E_1) = P(E_2)$ and $P(E_3) = P(E_4)$, then the FKG inequality of [32] yields

$$P(F) \geq P(\hat{\mathcal{D}}_{[n,\rho]}^r) P^2(E_1) P^2(E_3).$$

By lemma (43), there exists $b > 0$ such that for all sufficiently large M ,

$$P(F) \geq 1 - \frac{1}{2} M^{-c(r - r_0^*)} - O(e^{-b\phi(M)}).$$

Using $\phi(M) = o(c(r - r_0^*) \log M)$, this implies that for any given $\epsilon_1 > 0$ and all sufficiently large M depending upon ϵ_1 ,

$$P(F) \geq 1 - \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r - r_0^*)}.$$

It is now claimed that $F - \mathcal{D}_{[n,\rho-\delta]}^r = \emptyset$, requiring that $P(F - \mathcal{D}_{[n,\rho-\delta]}^r) = 0$ for all large M . Following [13], the occurrence of F implies that there is a connected path of hexagons which encloses the sub-lattice given by $H_{\mathcal{B}}(r) - \bigcup_{i=1}^4 H_i$. Because the points in \mathcal{X}_n are uniformly distributed, then there is a connected cluster of hexagons within the original lattice totaling at least $\rho M^2 - (2M\phi(M) + 2\phi(M)(M - 2\phi(M)))$ hexagons, where ρM^2 is a lower bound on the number of occupied hexagons in the largest connected cluster and $2M\phi(M) + 2\phi(M)(M - 2\phi(M))$ is the total number of hexagons in the strips, $H_i, i = 1, 2, 3, 4$. Let $\delta_1 = (2M\phi(M) + 2\phi(M)(M - 2\phi(M)))/M^2$. It follows that $F \subset \mathcal{D}_{[n,\rho-\delta_1]}^r$, since F occurs in those hexagons of \mathcal{B} that are not near the boundary of \mathcal{B} by a simple translation τ of hexagons $h \in \bigcup_{i=1}^4 H_i$ to hexagons $h \in H_{\mathcal{B}}(r) - \bigcup_{i=1}^4 H_i$. Thus, if M is large enough so that $\delta_1 < \delta$, then $F \subset \mathcal{D}_{[n,\rho-\delta_1]}^r \subset \mathcal{D}_{[n,\rho-\delta]}^r$. ■

Proof: (Theorem 33) Consider $r \leq r_0^*$. Since

$$P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) = P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}, \mathcal{D}_{[n,\rho]}^r) + P(\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r)$$

then

$$P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) \leq P(\mathcal{D}_{[n,\rho]}^r) + P(\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r).$$

It will be shown that $P(\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r) = o(M^{-c(r_0^* - r)})$. Let x be a configuration of states across hexagons in $H_{\mathcal{B}}(r)$ and let $\mathcal{C}(x) = \{C_1, \dots, C_K\}$ be the set of clusters in x . For $i = 1, \dots, K$, let N_{C_i} be the number of points in the cluster, C_i . Then, $\{N_{C_i} \mid \mathcal{C}(x), n\} \sim B(n, \frac{|H_{C_i}|}{M^2})$. Suppose $C_{i_0} \in \mathcal{C}(x)$ is any cluster such that $\rho_n(C_{i_0}) \geq \rho + \delta$. Since the occurrence of the property $(\mathcal{D}_{[n,\rho]}^r)^c$ implies $\frac{|H_{C_{i_0}}|}{M^2} < \rho$, then

$$\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r \subset \left\{ \rho_n(C_{i_0}) \geq \rho + \delta, \frac{|H_{C_{i_0}}|}{M^2} < \rho \right\}.$$

By arguments in [23] and [13], there is an $\alpha = \alpha(\rho, \delta) > 0$ such that

$$P\left(\rho_n(C_{i_0}) \geq \rho + \delta \mid \left\{ \frac{|H_{C_{i_0}}|}{M^2} < \rho \right\}, \mathcal{C}(x), n\right) \leq e^{-\alpha(\rho, \delta)n}.$$

$$\begin{aligned}
& \text{It follows that } P(\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r) \leq \\
& P\left(\left\{\rho_n(C_{i_0}) \geq \rho + \delta, \left\{\frac{|H_{C_{i_0}}|}{M^2} < \rho\right\}\right\}\right) \leq \\
& P\left(\rho_n(C_{i_0}) \geq \rho + \delta \mid \frac{|H_{C_{i_0}}|}{M^2} < \rho\right) \times P\left(\frac{|H_{C_{i_0}}|}{M^2} < \rho\right) \leq \\
& P\left(\rho_n(C_{i_0}) \geq \rho + \delta \mid \frac{|H_{C_{i_0}}|}{M^2} < \rho\right) = \\
& E\left[P\left(\rho_n(C_{i_0}) \geq \rho + \delta \mid \left\{\frac{|H_{C_{i_0}}|}{M^2} < \rho\right\}, \mathcal{C}(x), n\right)\right] \leq \\
& E[e^{-\alpha n}] = \exp\{-n(1 - e^{-\alpha})\}.
\end{aligned}$$

Now, since $n(1 - e^{-\alpha}) > d \log M$ implies $\exp\{-n(1 - e^{-\alpha})\} < M^{-d}$, then for any $d > 0$ and every fixed $\delta > 0$, it follows that $P(\mathcal{A}_{[n,\rho+\delta]}^{H_r} - \mathcal{D}_{[n,\rho]}^r)$ decays to zero at a rate faster than M^{-d} for n large enough. The case of $r \geq r_0^*$ is proven with similar arguments. ■

Theorem 44: $P(\mathcal{A}_{[n,\rho]}^{H_r})$ is a continuous function of ρ .

Proof: Let $\sigma = 1 - \rho$ in eq. (13). Then, $\mathcal{A}_{[n,\sigma]}^{H_r}$ is an increasing property in σ for increasing $\rho \in (\frac{1}{2}, 1)$. Therefore, by [33, Thm. (2.48)], it is true that $\mathcal{A}_{[n,\sigma]}^{H_r}$ has a sharp threshold in σ , and hence, in ρ . Thus, by [33, Ineq. (2.49)], $P(\mathcal{A}_{[n,\rho]}^{H_r})$ is differentiable in ρ , which implies that $P(\mathcal{A}_{[n,\rho]}^{H_r})$ is continuous as a function of ρ . ■

Remark 45: By thm. (44), for small $\delta > 0$,

$$P(\mathcal{A}_{[n,\rho-\delta]}^{H_r}) \approx P(\mathcal{A}_{[n,\rho]}^{H_r}) \approx P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}).$$

In this light, thm. (33) asserts that if $r_1^* < r_0^* < r_2^*$ and for some $\epsilon \in (0, \frac{1}{2})$, it is true that $P(\mathcal{A}_{[n,\rho]}^{H_{r_1^*}}) = \epsilon$ and $P(\mathcal{A}_{[n,\rho]}^{H_{r_2^*}}) = 1 - \epsilon$, then $r_2^* - r_1^*$ is an estimate of the sharp threshold interval length for the property, $\mathcal{A}_{[n,\rho]}^{H_r}$.

Proof: (Theorem 20) Since $P(\mathcal{A}_{[n,\rho]}^r)$ and $P(\mathcal{A}_{[n,\rho]}^{H_r})$ are continuous functions of r , then by thm. (33) and lemma (31), for every $r \in [0, r_0]$ there exists $r' \leq r$ such that

$$P(\mathcal{A}_{[n,\rho+\delta]}^{r'}) \leq P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) \quad (32)$$

$$\begin{aligned}
& \leq \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r_0^* - r)} \\
& \leq \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r_0 - r)}. \quad (33)
\end{aligned}$$

Consider $r_0 \in [0, r_0]$. Then, continuity of $P(\mathcal{A}_{[n,\rho]}^r)$ in r and the non-decreasing property of $P(\mathcal{A}_{[n,\rho]}^r)$ in r implies ineq. (33) for all $r \in [0, r']$. It is claimed that $r' = r_0$. Seeking a contradiction if $r' < r_0$, suppose $P(\mathcal{A}_{[n,\rho]}^r) \leq (\frac{1}{2} + \epsilon_1) M^{-c(r_0 - r)}$ for all $r \in [0, r']$ and $P(\mathcal{A}_{[n,\rho]}^r) > (\frac{1}{2} + \epsilon_1) M^{-c(r_0 - r)}$ for all $r > r'$. By hypothesis, $r_0 > r'$ so that when $r = r_0$, it follows that $P(\mathcal{A}_{[n,\rho+\delta]}^{r_0}) > 1/2$. Now, since for any connected cluster $\langle C \rangle_r$ such that $\rho_n(C) \geq \rho + \delta$ for $\delta > 0$, the statement $\rho_n(C) \geq \rho$ is implied, then $\mathcal{A}_{[n,\rho+\delta]}^r \subseteq \mathcal{A}_{[n,\rho]}^r$ for all $r \in [0, r_0]$. Hence, $r' < r_0$ leads to

$$P(\mathcal{A}_{[n,\rho]}^{r_0}) \geq \limsup_{\delta \rightarrow 0^+} P(\mathcal{A}_{[n,\rho+\delta]}^{r_0}) \geq P(\mathcal{A}_{[n,\rho+\delta]}^{r_0}) > \frac{1}{2}. \quad (34)$$

In particular, ineq. (34) gives $P(\mathcal{A}_{[n,\rho]}^{r_0}) > 1/2$. This is a contradiction since $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1/2$ by thm. (14). It follows

that $r' = r_0$ and

$$P(\mathcal{A}_{[n,\rho]}^r) \leq \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r_0 - r)}$$

for $r \leq r_0$. A similar argument is used to prove

$$P(\mathcal{A}_{[n,\rho-\delta]}^r) \geq 1 - \left(\frac{1}{2} + \epsilon_1\right) M^{-c(r - r_0)}$$

for $r \geq r_0$. ■

The implication of the proof to thm. (20) is that $P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}_{[n,\rho]}^{H_r})$ for $r \in [0, r_0]$. By [33, Thm. (1.16)], the random cluster measure gives rise to a collection of conditional probability measures of connection events in the identified classes during K -means classification. Therefore, the node process X samples from each element of the collection.

Theorem 46: $P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}_{[n,\rho]}^{H_r})$ for $r \in [0, r_0]$.

Proof: By continuity in ρ of $P(\mathcal{A}_{[n,\rho]}^{H_r})$ as given by thm. (44), it is true that

$$\lim_{\delta \rightarrow 0^+} P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) = P(\mathcal{A}_{[n,\rho]}^{H_r}).$$

Suppose $\delta_1 > \delta_2$ such that $\rho + \delta_1, \rho + \delta_2 \in (\frac{1}{2}, 1)$ and let $\langle C \rangle_r \in \mathcal{A}_{[n,\rho+\delta_1]}^r$. Then, $\rho_n(C) \geq \rho + \delta_1 > \rho + \delta_2$ so that $\langle C \rangle_r \in \mathcal{A}_{[n,\rho+\delta_2]}^r$. Hence, $\mathcal{A}_{[n,\rho+\delta_1]}^r \subseteq \mathcal{A}_{[n,\rho+\delta_2]}^r$. By properties of probability measures, $P(\mathcal{A}_{[n,\rho]}^r)$ is monotone non-decreasing as a function of decreasing ρ . By ineq. (32), it follows that for some fixed $r \in [0, r_0]$, there exists $r' \in [0, r_0]$ such that $P(\mathcal{A}_{[n,\rho+\delta]}^{r'}) \leq P(\mathcal{A}_{[n,\rho+\delta]}^{H_r})$ for all $r'' \in [0, r']$ so that

$$\limsup_{\delta \rightarrow 0^+} P(\mathcal{A}_{[n,\rho+\delta]}^{r''}) \leq \limsup_{\delta \rightarrow 0^+} P(\mathcal{A}_{[n,\rho+\delta]}^{H_r}) = P(\mathcal{A}_{[n,\rho]}^{H_r}). \quad (35)$$

From the proof of thm. (20), it was shown that $r' = r_0$. Therefore, by continuity of $P(\mathcal{A}_{[n,\rho]}^{H_r})$ in r , ineq. (35) holds for all $r \in [0, r_0]$, with r'' replaced by r . The Monotone Convergence Theorem [67] applied to $E[1_{\mathcal{A}_{[n,\rho+\delta]}^r}]$ and $E[1_{\mathcal{A}_{[n,\rho]}^r}]$ guarantees that $P(\mathcal{A}_{[n,\rho+\delta]}^r) \rightarrow P(\mathcal{A}_{[n,\rho]}^r)$ as $\delta \rightarrow 0^+$. Therefore, ineq. (35) becomes

$$P(\mathcal{A}_{[n,\rho]}^r) = \limsup_{\delta \rightarrow 0^+} P(\mathcal{A}_{[n,\rho+\delta]}^r) \leq P(\mathcal{A}_{[n,\rho]}^{H_r}). \quad (36)$$

In particular, $P(\mathcal{A}_{[n,\rho]}^r) \leq P(\mathcal{A}_{[n,\rho]}^{H_r})$ so that with the result of lemma (30), namely $P(\mathcal{A}_{[n,\rho]}^{H_r}) \leq P(\mathcal{A}_{[n,\rho]}^r)$, the theorem follows. ■

Corollary 47: $P(\mathcal{A}^r) = P(\mathcal{A}^{H_r})$ for $r \in [0, r_0]$.

Proof: By thm. (46), it is true that $P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}_{[n,\rho]}^{H_r})$ for all $r \in [0, r_0]$ and all $n \geq 1$. By prop. (77), it follows that $P(\mathcal{A}^{H_r}) \leq P(\mathcal{A}_{[n,\rho]}^{H_r}) = P(\mathcal{A}_{[n,\rho]}^r)$. In particular, $P(\mathcal{A}^{H_r}) \leq P(\mathcal{A}_{[n,\rho]}^r)$. Without loss of generality, assume that $\text{area}(\mathcal{B}) = 1$. From [13], differentiability of $P(\mathcal{A}_{[n,\rho]}^r)$ in point density $\lambda = \lambda(n) = E[n]$ implies continuity of $P(\mathcal{A}_{[n,\rho]}^r)$ in λ so that the following holds

$$\lim_{E[n] \rightarrow \infty} P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}^r). \quad (37)$$

Therefore, $P(\mathcal{A}^{H_r}) \leq P(\mathcal{A}_{[n,\rho]}^r)$ and eq. (37) implies $P(\mathcal{A}^{H_r}) \leq P(\mathcal{A}^r)$. Similarly, $P(\mathcal{A}^r) \leq P(\mathcal{A}^{H_r})$ so that the corollary follows. ■

Corollary 48: $r_0 = r_0^*$.

Proof: By thm. (46), it is true that $1/2 = P(\mathcal{A}_{[n,\rho]}^{r_0}) = P(\mathcal{A}_{[n,\rho]}^{H_{r_0}})$. In particular, $1/2 = P(\mathcal{A}_{[n,\rho]}^{H_{r_0^*}})$. Since $P(\mathcal{A}_{[n,\rho]}^{H_{r_0^*}}) = 1/2 = P(\mathcal{A}_{[n,\rho]}^{H_{r_0}})$, by the discussion preceding thm. (32) and by thm. (14), then the uniqueness of r_0^* and r_0 guarantees that $r_0^* = r_0$. ■

Proof: (Theorem 21) Follows directly from thms. (44) and (46). ■

By thm. (46) and cor. (48), the problem of estimating the probabilities and length of the sharp threshold interval in the continuum can be re-cast as problems of estimation in the presence of a hexagonal partition of the bounded region. As such, tools from percolation [32] and the random cluster model [33] can readily be employed. This fact will be of paramount importance in applications to K -means classification where a data set consisting of multi-dimensional points is partitioned into disjoint, connected subsets. As it is advantageous to not have one connected cluster containing at least $100\rho\%$ of all points, since otherwise there may exist a single cluster containing almost all points by lemma (72), the connection radius for points in the continuum must be in the sub-critical range $r \in [0, r_0]$ when classifying data into more than 2 classes. Since $P(\mathcal{A}_{[n,\rho]}^r) = P(\mathcal{A}_{[n,\rho]}^{H_r})$ for $r \in [0, r_0]$, disjoint clusters of points in the continuum are equivalent to disjoint clusters of occupied hexagons in the hexagonal partition of the bounded region containing all points. As such, multi-dimensional points in the continuum can be thought to belong to the same class if they are within a certain Euclidean distance of one another. As a result, the multi-dimensional points will have representatives belonging to occupied, connected hexagons in the 2-dimensional, bounded, partitioned region. All representatives in connected clusters of hexagons form the members of a class.

IV. K -MEANS SHARP THRESHOLD AND CRITICAL RADIUS

For ease of computation, and at the risk of ambiguity, suppose $n = M^2$. The idea is to partition \mathcal{B} into M^2 hexagons and find $K = N^2$ contiguous clusters of hexagons such that each of the clusters are mutually disjoint. Into one and only one hexagon of a given cluster will each data point be mapped to form a point in the connected cluster. As such, the connected clusters of hexagons will be the $K = N^2$ classes containing a representative point associated to one and only one data point.

Theorem 49: Assume that there are M^2 points and N^2 classifications for the points. The minimum number of hexagons required to partition the unit square into N^2 disjoint regions such that M^2 is the sum total of all hexagons in the disjoint regions is given by

$$S(M, N) = M^2 + 2M(N - 1)^2.$$

Proof: Since $M^2 \gg N^2$ by hypothesis, then the total number of hexagons required to partition \mathcal{B} into disjoint regions of contiguous hexagons is $O(M^2)$. Label the disjoint regions A_1, A_2, \dots, A_{N^2} and let k be any integer such that $1 \leq k \leq N^2$. Since the total number of hexagons partitioning \mathcal{B} is $O(M^2)$, then the number of hexagons in A_k is proportional to

M^2 . Likewise, the total number of hexagons in boundary(A_k) is proportional to $\text{area}(A_k)$. Since $\text{area}(A_k)$ is proportional to M^2 , then the number of hexagons in boundary(A_k) is proportional to M^2 . Note that each A_k shares a portion of its separating boundary with each of its neighboring clusters of hexagons. Let A_j be a neighboring cluster of A_k such that $j \neq k$ and $1 \leq j \leq N^2$. Since this portion of the separating boundary is proportional to both $\text{area}(A_k)$ and $\text{area}(A_j)$, then it is proportional to a common area of size $\text{area}(A_{kj})$. Repeating this same logic for all integers k and j such that $1 \leq k \leq N^2$ and $1 \leq j \leq N^2$, the total number of hexagons in the entire separating boundaries is proportional to a common area of size $\text{area}(A)$. Since minimizing the total number of hexagons in \mathcal{B} is tantamount to minimizing the $\text{area}(A)$, then making an application of the law of large numbers, each of the N^2 disjoint clusters of connected hexagons is the same size and must be a square sub-region of \mathcal{B} containing M^2/N^2 hexagons. The minimum number of hexagons that are required to enclose N^2 sub-regions of \mathcal{B} containing M^2/N^2 hexagons is exactly $2M(N - 1)^2$. Therefore, the minimum number of hexagons required to partition \mathcal{B} into N^2 disjoint regions such that M^2 is the sum total of all hexagons in the disjoint regions is given by

$$S(M, N) = M^2 + 2M(N - 1)^2. \quad (38)$$

The idea is to use the result of the theorem to calculate, as a function of M and $N = N(M)$, the exact size of a prototypical hexagon which will be used to partition \mathcal{B} into hexagons of equal size. As $K = N^2$ is fixed as the number of classes of data points, M^2 is fixed for the initial calculation of $S(M, N)$ and the subsequent classification of the first M^2 data points. In [32], it is stated and proven that there is a critical probability of connection between hexagons containing a point of a network such that it is no longer possible to have disjoint clusters of points when this critical probability of connection is exceeded. Hence, all points will be connected into one cluster, which is not what we intend to model, in this case. Since the size of \mathcal{B} is fixed, then to decrease the probability of connection while maintaining $K = N^2$ disjoint contiguous clusters of points, the size of each hexagon must decrease while increasing the number of hexagons in the boundaries of the disjoint regions. In this way, the ratio of the total number of occupied hexagons to the total number of hexagons will be less than this critical probability of connection. Note that we used uniformity of the points throughout \mathcal{B} so that the approximate number of points in a cluster of hexagons is proportional to the ratio of the number of hexagons in the cluster divided by the number of hexagons in the entire region, \mathcal{B} . Also, note that the minimum number of hexagons required for separation is given by thm. (49), so that the common radius of the circle that can circumscribe any one of these hexagons is of size

$$R(M, N) = \frac{1}{2\sqrt{S(M, N)}}, \quad (39)$$

thereby necessarily indicating that

$$B(M, N) = 2 * R(M, N)$$

is the diameter of the circumscribing circle. $R(M, N)$ is decreasing for increasing M and N as a direct result of eqs. (38) and (39).

Lemma 50: $R(M, N)$ is decreasing for increasing M and N .

Proof: By eq. (38), $S(M, N)$ is increasing for increasing M and N . Consequently, by eq. (39), $R(M, N)$ is decreasing for increasing M and N . ■

Theorem 51: Suppose that the node process X generates infinitely many points in \mathbb{R}^2 . An infinite connected cluster exists across hexagons in \mathbb{R}^2 with probability 1 if and only if the probability that any two points connect exceeds p_c , where p_c is the critical probability of connection. Otherwise, all connected clusters are disjoint with probability 1.

Theorem (51) is a restatement of [33, *Thm.* (1.11)]. A direct result of thm. (51) is that, given any bounded region \mathcal{B} , all points generated within \mathcal{B} are almost surely connected into one cluster. Therefore, in order to not exceed the critical probability of connection, which means maintaining the N^2 classes of M^2 data points, the radial length of each hexagon's circumscribing circle must be less than or equal to $R(M, N)$. By [32, *Thm.* (1.11)], the clusters will be disjoint with probability 1. Hence, the following corollary to thm. (49) follows from these statements and lemma (53).

Corollary 52: Let h^r be a hexagon of size such that it can be inscribed into a circle of radius $r = r(M, N) > 0$ where

$$0 < r \leq R(M, N).$$

If \mathcal{B} is partitioned into copies of h^r , then with probability 1, N^2 is the mean number of disjoint clusters of contiguous hexagons in the region \mathcal{B} that are occupied by the M^2 points.

With r_0 given by cor. (52), the size of the prototypical hexagon can be calculated for repartitioning \mathcal{B} . Furthermore, cor. (52) guarantees that the classes will remain distinct, with probability 1, through each new classification. By cor. (52), the expected value of the number of classes to form can be calculated.

Lemma 53: For M^2 uniformly distributed data points in \mathcal{B} and for any $\rho \in (0, p_c]$, with $p_c = 1 - 2 \sin(\pi/18)$,

$$\frac{M^2}{S(M, N)} = \frac{M^2}{M^2 + 2M(N-1)^2} = \rho \quad (40)$$

determines the expected number $K = N^2$ of disjoint classes to form such that M^2 is the total of all occupied hexagons across all classes.

Proof: At the risk of ambiguity, let N^2 denote both the random variable and the expectation of the random variable which takes the number of formed classes as its value. Because \mathcal{B} is partitioned by hexagons, it is shown in [33, *Chapter 3*] that $p_c = 1 - 2 \sin(\pi/18)$. By uniformity, the mean number of data points in each class is M^2/N^2 . By thm. (51), each class will be disjoint and each hexagon in \mathcal{B} will be as large as possible if \mathcal{B} is partitioned into $S(M, N)$ hexagons of equal size. Also, by thm. (51), the probability of any of the M^2 hexagons being populated with a data point has to be less than or equal to p_c in order that the expected classes form with probability 1, resulting in eq. (40). For any $\rho \in (0, p_c]$, $K = N^2$ is found by solving eq. (40) to obtain $K = N^2$ as the least integer which is not less than the integer part of a non-negative solution to eq. (40), for fixed, positive M^2 . ■

Lemma 54: For fixed $\rho \in (\frac{1}{2}, 1)$ and $r > 0$ there exists $\delta = \delta(\rho) \in (0, \frac{1}{2})$, such that

$$\left\{ \frac{|\langle C \rangle_{H_r}|}{S(M, N)} < \frac{1}{2} \right\} = \left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$$

upto sets of P -measure zero.

Proof: By definition, $\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c = \left\{ \frac{|\langle C \rangle_{H_r}|}{S(M, N)} < \rho - \delta \right\}$. Take $\delta = \rho - \frac{1}{2}$. ■

By choosing δ as in lemma (54), continuity in $r > 0$ and the non-decreasing property of $P \left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)$ for increasing $r > 0$ granted by cor. (12) and prop. (76), respectively, then by ineq. (10), it follows that

$$R(M, N) < r_0^* = r_0^*(M, N)$$

for the property $\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$, since

$$\begin{aligned} P \left(\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_{R(M, N)}} \right)^c \right) &= 1 \\ &> \frac{1}{2} \\ &= P \left(\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_{r_0^*}} \right)^c \right) \end{aligned}$$

and the probability of $\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$ is non-decreasing for decreasing $r \leq r_0^*$, a reversal.

Let $\epsilon \in (0, \frac{1}{2})$ be given and let $r_1^* > 0$ and $r_2^* > 0$, guaranteed by cor. (12), be such that $P \left(\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_{r_1^*}} \right)^c \right) = 1 - \epsilon$ and $P \left(\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_{r_2^*}} \right)^c \right) = \epsilon$, respectively. Then, again by cor. (12), it follows that

$$R(M, N) < r_1^* < r_0^* = r_0^*(M, N) < r_2^*.$$

By symmetry, it follows that

$$R(M, N) < r_1^* < r_0^* = r_0^*(M, N) < r_2^* < 2r_0^* - R(M, N). \quad (41)$$

Note that by cor. (52) and by symmetry,

$$P \left(\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c \right) = 0$$

when $r \geq 2r_0^* - R(M, N)$. Therefore, if $\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$ occurs with probability 0, then the property $\left\{ \frac{M^2}{S(M, N)} < \frac{1}{2} \right\}$ occurs with probability 0. Otherwise, $\left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$ would occur with positive probability, since $\left\{ \frac{M^2}{S(M, N)} < \frac{1}{2} \right\} \subseteq \left\{ \frac{|\langle C \rangle_{H_r}|}{S(M, N)} < \frac{1}{2} \right\} = \left(\mathcal{A}_{[S(M, N), \rho - \delta]}^{H_r} \right)^c$, upto sets of P -measure zero, by lemma (54). Hence, $\left\{ \frac{M^2}{S(M, N)} \geq \frac{1}{2} \right\}$ occurs with probability 1. As a result,

$$\frac{M^2}{M^2 + 2M(N-1)^2} \geq \frac{1}{2} \quad (42)$$

with probability 1. Therefore, with probability 1 for M , it follows that N is a solution to $M^2 + 2M(N-1)^2 - 2M^2 = 0$.

Lemma 55: If $r \geq 1/(2N)$, then $P\left(\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^{H_r}\right)^c\right) = 0$.

Proof: Without loss of generality, suppose $\text{area}(\mathcal{B}) = 1$ and further suppose that \mathcal{B} is divided into squares with sides of length $2r = 1/N$. By hypothesis, \mathcal{B} contains M^2 data points and it is to be divided into N^2 regions. Clearly then, there are no boundary hexagons separating each of the N^2 regions since the sides of \mathcal{B} have length $2rN = 1$ which gives \mathcal{B} an area of 1. Let each square be inscribed by a circle of radius r , which itself is inscribed by a hexagon. By hypothesis, each of the N^2 regions in \mathcal{B} contains at least one of the M^2 data points. Hence, each of the N^2 (occupied) regions is connected in a cluster to every other region in \mathcal{B} so that $P\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^{H_r}\right) = 1$. Since $P\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^{H_r}\right) = 1$ for $r = 1/(2N)$, then $P\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^{H_r}\right) = 1$ for $r \geq 1/(2N)$ by prop. (76). ■

As a result of lemma (55) and by using ineq. (41), a conservative estimate for r_0^* is given by a solution to

$$2r_0^* - R(M, N) \geq \frac{1}{2N} \quad (43)$$

that maximizes $1/(2N)$ as a function of M . The value of N satisfies ineq. (42) and a maximal solution is found when equality holds. As such, for $\epsilon \in (0, \frac{1}{2})$, since $(r_1^*, r_2^*) \subset (R(M, N), 2r_0^* - R(M, N))$, then by ineq. (43),

$$\begin{aligned} r_2^* - r_1^* &\approx 2r_0^* - 2R(M, N) \\ &= \frac{1}{2N} - R(M, N) \end{aligned} \quad (44)$$

is an estimate of the length of the sharp threshold interval $r_2^* - r_1^*$ about r_0^* .

Using the value of r_0^* given by eq. (43) and by using the estimate for the length of the sharp threshold interval about r_0^* given by eq. (44), an estimate for the value of r_1^* can be obtained. Thus, when $r \leq r_1^*$, the property $\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^{H_r}\right)^c$ occurs with probability at least $1 - \epsilon$ and falls sharply to a probability of occurrence of ϵ as $r \rightarrow r_2^*$.

By cor. (48) and thm. (46), the left half of the sharp threshold interval about r_0 is given by $[r_1^*, r_0]$. Using lemma (30), there exists $r_2 \leq r_2^*$ such that $[r_0, r_2]$ is the right half of the sharp threshold interval for $\epsilon > 0$ given. Thus, when $r \leq r_1^*$, the property $\left(\mathcal{A}_{[S(M,N),\rho-\delta]}^r\right)^c$ occurs with probability at least $1 - \epsilon$ and falls sharply to a probability of occurrence of (no greater than) ϵ as $r \rightarrow r_2^*$. As such, the sharp threshold interval for classifying M^2 data points into N^2 classes, in the mean continuum case, is of length (no greater than) $r_2^* - r_1^*$.

Theorem 56: Let $\Delta^*(M, N)$ denote the sharp threshold interval length for the event of classifying M^2 random data points into N^2 classes. Then,

$$\Delta^*(M, N) = O(N^{-1}).$$

Proof: Follows directly from eq. (44), eq. (39) and thm. (49). ■

V. THE NEURAL NETWORK

Let T^2 be the total number of rows in the data set from which the M^2 samples are taken. From lemma (50), recall that $R(M, N)$ is decreasing for increasing M . As such, once the actual number of classes to form, N_0^2 , is known, then given $M^2 \ll T^2$, for classifications of the ordered data set, it follows that $R(M, N_0)$ constitutes the actual upper bound on the distance that any class member is allowed to differ from the center of its respective class. As such, any anomalies, which by definition lie outside of any regularized class, will be at a distance greater than $R(M, N_0)$ from the regression hyperplane formed from the union of the regularized classes. This distance is measured as the length of the projection of an anomaly onto the normal vector of the regression hyperplane. By normality, the center of the union of regularized classes is the maximum likelihood estimate (MLE) of the union, which defines the best linear estimate, given the data. Thus, the regression hyperplane necessarily passes through the center of the union formed after the K -means clustering process.

Since the initial classification of the ordered set was done with $R(M, N)$ as the upper bound, then class members are candidate anomalies when they fall outside the new upper bound $R(M, N_0)$ in distance from the regression hyperplane. Furthermore, the same restriction is applied to the class centers, whereby, given that the regression hyperplane is the linear average of the data points used in its definition, then anomalous classes, those sparsely populated clusters of points which lie outside of the union of the regularized classes, are those with centers that deviate from the regression hyperplane by more than a distance of $R(M, N_0)$.

A. Definitions

Definition 57: (Macro Anomaly Detection) Let N_0^2 be the actual number of classes to form after classification of all data points has completed. Suppose the regression hyperplane $H_{[M, N_0]}^\theta$ is given by $(w_{[M, N_0]}^\theta)^t z = \theta$, for some real vector $w_{[M, N_0]}^\theta$ and some constant $\theta \in \mathbb{R}$. Let $h \in \{1, \dots, N_0^2\}$ and $H_{[M, N_0]}^\theta$ denote the regression hyperplane formed from the union of the regularized classes. Then, a class C_h is *anomalous*, if $x_{(h)}$ is the center of C_h and $d(x_{(h)}, H_{[M, N_0]}^\theta) \geq R(M, N_0)$. Otherwise, C_h is *NON-anomalous*.

Definition 58: (Micro Anomaly Detection) Given $h \in \{1, \dots, N_0^2\}$, a data point $x \in C_h$ is *anomalous*, if $d(x, H_{[M, N_0]}^\theta) \geq R(M, N_0)$ and *NON-anomalous* otherwise.

B. Anomaly Segregation and The Decision Boundary

The combined definitions of macro and micro anomaly detection given in defs. (57) and (58) simply states that the non-anomalous data should all be tightly wrapped in the interior of hyperspheres of diameter no more than $2 * R(M, N_0)$, with each of the weighted centers being of distance no more than $R(M, N_0)$ from the regression hyperplane, once classification has ceased.

1) Regularity Characterized:

Lemma 59: Given $h \in \{1, \dots, N_0^2\}$, a class C_h is *NON-anomalous* if and only if $d(x_{(h)}, H_{[M, N_0]}^\theta) < R(M, N_0)$.

Furthermore, if C_h is NON-anomalous, then $d(x, y) < 2 * R(M, N_0)$ for all $x, y \in C_h$.

Proof: The first part follows from the definition of an anomalous class given in section (57). For the second part, the triangle inequality and the above definitions of anomalous class and anomalous data point in sections (57) and (58), respectively, are used to obtain

$$\begin{aligned} d(x, y) &\leq d(x, H_{[M, N_0]}^\theta) + d(H_{[M, N_0]}^\theta, y) \\ &< R(M, N_0) + R(M, N_0) \\ &= 2 * R(M, N_0). \end{aligned}$$

■

2) An Anomaly Segregation Theorem:

Theorem 60: Given $h \in \{1, \dots, N_0^2\}$, if class C_h is anomalous, then there exists at least one anomalous data point, $x \in C_h$.

Proof: Let $H_{[M, N_0]}^\theta$ be defined as before and define $X_{[M, N_0]}^{(h)} = \{x \in C_h \mid d(x, H_{[M, N_0]}^\theta) \geq R(M, N_0)\}$. Seeking a contradiction, suppose $X_{[M, N_0]}^{(h)} = \emptyset$. Then, for every $x \in C_h$, it is true that $d(x, H_{[M, N_0]}^\theta) < R(M, N_0)$. Therefore, the contradiction is obtained and the theorem is proven, if it can be shown that there exists $x \in C_h$ such that $d(x_{(h)}, H_{[M, N_0]}^\theta) \leq d(x, H_{[M, N_0]}^\theta)$, which implies $d(x_{(h)}, H_{[M, N_0]}^\theta) < R(M, N_0)$, the sought contradiction to class C_h being anomalous. Thus, if $|C_h| = 1$, then $x_{(h)} = x \in C_h$. Otherwise, suppose that $|C_h| > 1$ and assume that $d(x_{(h)}, H_{[M, N_0]}^\theta) > d(x, H_{[M, N_0]}^\theta)$ for all $x \in C_h$. Let v_h be the vector normal to $H_{[M, N_0]}^\theta$ which passes through $x_{(h)}$ and let v_h^t denote its transpose. If $\hat{x} \in C_h$ is such that $d(\hat{x}, H_{[M, N_0]}^\theta) \geq d(x, H_{[M, N_0]}^\theta)$ for all $x \in C_h$, then

$$\begin{aligned} d(x_{(h)}, H_{[M, N_0]}^\theta) &= \frac{|v_h^t x_{(h)}|}{\|v_h\|} \\ &= \frac{|v_h^t \frac{\sum_{x \in C_h} x}{|C_h|}|}{\|v_h\|} \\ &= \|v_h\|^{-1} \frac{|\sum_{x \in C_h} v_h^t x|}{|C_h|} \end{aligned} \quad (45)$$

and

$$\begin{aligned} d(\hat{x}, H_{[M, N_0]}^\theta) &= \frac{|v_h^t \hat{x}|}{\|v_h\|} \\ &= \|v_h\|^{-1} \frac{|\sum_{x \in C_h} v_h^t \hat{x}|}{|C_h|} \end{aligned} \quad (46)$$

so that eq. (46), together with $d(x_{(h)}, H_{[M, N_0]}^\theta) > d(x, H_{[M, N_0]}^\theta)$ and the triangle inequality applied to eq. (45) implies $d(x, H_{[M, N_0]}^\theta) = \|v_h\|^{-1} |v_h^t x| > \|v_h\|^{-1} |v_h^t \hat{x}| = d(\hat{x}, H_{[M, N_0]}^\theta)$ for all $x \in C_h$. This contradicts $d(\hat{x}, H_{[M, N_0]}^\theta) \geq d(x, H_{[M, N_0]}^\theta)$ for all $x \in C_h$. Thus, $d(x_{(h)}, H_{[M, N_0]}^\theta) \leq d(x, H_{[M, N_0]}^\theta)$ for at least one $x \in C_h$, which is the originally-sought contradiction. ■

Theorem (60) provides a means for segregating all anomalous data points from designated anomalous classes, leaving only classes consisting of non-anomalous data points. For each

$h \in \{1, \dots, N_0^2\}$, let $X_{[M, N_0]}^{(h)} = \{x \in C_h \mid d(x, H_{[M, N_0]}^\theta) \geq R(M, N_0)\}$, as in the proof of thm. (60). Then, $(X_{[M, N_0]}^{(h)})^c = \{x \in C_h \mid d(x, H_{[M, N_0]}^\theta) < R(M, N_0)\}$ is a class of non-anomalous data points for each $h \in \{1, \dots, N_0^2\}$. Define

$$X_{[M, N_0]} = \bigcup_{h=1}^{N_0^2} X_{[M, N_0]}^{(h)}.$$

Definition 61: $\Omega = X_{[M, N_0]}^c \cup X_{[M, N_0]}$ is pointwise linearly separable, if there exists $x \in \Omega$ and a subset $A_x \subset \Omega$ such that $w_x^t y \leq \theta$ for all $y \in A_x$ and $w_x^t y > \theta$ for all $y \in \Omega \setminus A_x$, where $w_x \in \mathbb{R}^L$, $L \geq 2$ and $\theta \in \mathbb{R}$.

Theorem 62: If $X_{[M, N_0]} \neq \emptyset$, then Ω is pointwise linearly separable into $X_{[M, N_0]}$ and $X_{[M, N_0]}^c$ for some $x \in X_{[M, N_0]} \subset \Omega$.

Proof: For some specific $h \in \{1, \dots, N_0^2\}$ to be chosen later, suppose $x \in X_{[M, N_0]}^{(h)} \subset X_{[M, N_0]}$. The idea is to shift $H_{[M, N_0]}^\theta$ by a certain length along the vector normal to $H_{[M, N_0]}^\theta$ which passes through $x_{(h)}$. Thus, without loss of generality, suppose x lies to one side of $H_{[M, N_0]}^\theta$ so that $(w_{[M, N_0]}^\theta)^t x \leq \theta$ is a hyper half-plane. Define $\hat{y} \in X_{[M, N_0]}^c$ to be a vector such that $d(\hat{y}, H_{[M, N_0]}^\theta) \geq d(y, H_{[M, N_0]}^\theta)$ for all $y \in X_{[M, N_0]}^c$, where $d(y, H_{[M, N_0]}^\theta)$ is the length of the projection of y onto the vector normal to $H_{[M, N_0]}^\theta$, and is given by

$$d(y, H_{[M, N_0]}^\theta) = \|w_{[M, N_0]}^\theta\|^{-1} (w_{[M, N_0]}^\theta)^t y.$$

Define

$$d_{[M, N_0]}^\theta = \min_{x \in X_{[M, N_0]}^{(h)}} \left(d(x, H_{[M, N_0]}^\theta) - d(\hat{y}, H_{[M, N_0]}^\theta) \right). \quad (47)$$

Now, $h \in \{1, \dots, N_0^2\}$ can be chosen such that $x \in X_{[M, N_0]}^{(h)}$ is a vector which satisfies eq. (47). For $\gamma \in (0, 1)$, let $w_{[M, N_0]}^{\theta_\gamma}$ be a real vector such that $(w_{[M, N_0]}^{\theta_\gamma})^t z = \theta - \theta_\gamma$ is the hyperplane

$$H_{[M, N_0]}^{\theta_\gamma} = H_{[M, N_0]}^\theta + \left(\theta_\gamma \times \frac{w_{[M, N_0]}^\theta}{\|w_{[M, N_0]}^\theta\|} \right), \quad (48)$$

where

$$\theta_\gamma = d(x, H_{[M, N_0]}^\theta) - \gamma d_{[M, N_0]}^\theta, \quad (49)$$

with the right side of eq. (48) being a vector sum for each vector in $H_{[M, N_0]}^\theta$. Thus, by copying and shifting the regression hyperplane $H_{[M, N_0]}^\theta$ along the direction of its normal vector in order to obtain $H_{[M, N_0]}^{\theta_\gamma}$ and by considering the reflection of the shifted hyperplane across the regression hyperplane, it now follows that

$$(w_{[M, N_0]}^{\theta_\gamma})^t y \leq (\theta + \theta_\gamma), \quad (50)$$

for all $y \in X_{[M, N_0]}$ such that $(w_{[M, N_0]}^\theta)^t y \leq \theta$ and

$$(w_{[M, N_0]}^{\theta_\gamma})^t y > (\theta + \theta_\gamma) \quad (51)$$

for all $y \in X_{[M,N_0]}$ such that $\left(w_{[M,N_0]}^\theta\right)^t y > \theta$, with the opposite of inequalities (50) and (51) otherwise, for $y \in X_{[M,N_0]}^c$. Given $M^2, T^2, N_0^2, \gamma \in (0, 1)$ and some fixed $\theta \in \mathbb{R}$, take $w_x = w_{[M,N_0]}^{\theta^x}$ for $x \in X_{[M,N_0]}$ which satisfies eq. (47). ■

Theorem (62) provides the means for identifying the decision boundary to be used when determining if certain data points are anomalous, with ineq. (50) or (51) defining the shifted regression hyperplane.

3) *The Neural Network Anomaly Detector:* By thm. (62), with the anomalous data points segregated and collected into the set $X_{[M,N_0]}$, it's now possible to store the anomaly detector offline as the set of synaptic weights of a two-class discriminating neural network, which can be designed as a perceptron with a single input layer used to compute the synaptic weights $w_{[M,N_0]}^{\theta^x}$ associated with copying and shifting the regression hyperplane $H_{[M,N_0]}^\theta$, given by $\left(w_{[M,N_0]}^\theta\right)^t z = \theta$, in the direction of the normal vector to $H_{[M,N_0]}^\theta$, for some $x \in X_{[M,N_0]}^{(h)}$ and $h \in \{1, \dots, N_0^2\}$.

Theorem 63: (Neural Network Anomaly Detector) Suppose $X_{[M,N_0]} \neq \emptyset$ and let $x \in X_{[M,N_0]}$ satisfy eq. (47), with $w_{[M,N_0]}^{\theta^x}$ defined by

$$w_{[M,N_0]}^{\theta^x} = w_{[M,N_0]}^\theta + \left(\theta_\gamma^x \times \frac{w_{[M,N_0]}^\theta}{\|w_{[M,N_0]}^\theta\|} \right), \quad (52)$$

for some chosen $\gamma \in (0, 1)$. For all newly sampled data points $y \in \Omega$, define $\phi_{[M,N_0]}^{\theta^x} : \Omega \rightarrow \mathbb{R}$ as

$$\phi_{[M,N_0]}^{\theta^x}(y) = \left(w_{[M,N_0]}^{\theta^x}\right)^t y - (\theta + \theta_\gamma^x). \quad (53)$$

Then, the activation function $\phi_{[M,N_0]}^{\theta^x}$, along with the synaptic weight vector $w_{[M,N_0]}^{\theta^x}$, defines a two-class discriminating neural network such that $y \in \Omega$ is anomalous if for some $\hat{\theta}_\gamma^x \in \mathbb{R}$, the reflection of $\phi_{[M,N_0]}^{\theta^x}$ across $w_{[M,N_0]}^\theta = \theta$, given by $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}$, satisfies $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}(y) > 0$ whenever $\left(w_{[M,N_0]}^\theta\right)^t y \leq \theta$ or if $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}(y) \leq 0$ whenever $\left(w_{[M,N_0]}^\theta\right)^t y > \theta$. Otherwise, $y \in \Omega$ is non-anomalous.

Proof: The synaptic weight vector $w_{[M,N_0]}^{\theta^x}$ given in eq. (52) follows since $H_{[M,N_0]}^{\theta^x}$ given in eq. (48) is uniquely determined by shifting the vector $w_{[M,N_0]}^\theta$, normal to $H_{[M,N_0]}^\theta$ at the origin, in the direction which is determined by ineq. (50). Without loss of generality, suppose $\theta = 0$ and further suppose $\left(w_{[M,N_0]}^\theta\right)^t y \leq \theta$. The neural network anomaly detector,

given by the activation function $\phi_{[M,N_0]}^{\theta^x}$ in eq. (53), gives

$$\begin{aligned} \phi_{[M,N_0]}^{\theta^x}(y) &= \left(w_{[M,N_0]}^{\theta^x}\right)^t y - (\theta + \theta_\gamma^x) \\ &= \left(w_{[M,N_0]}^\theta\right)^t y \\ &\quad + \left(\theta_\gamma^x \times \frac{\left(w_{[M,N_0]}^\theta\right)^t y}{\|w_{[M,N_0]}^\theta\|} \right) - (\theta + \theta_\gamma^x) \end{aligned} \quad (54)$$

$$\begin{aligned} &= \left(w_{[M,N_0]}^\theta\right)^t y - \theta \\ &\quad + \left(\theta_\gamma^x \times \frac{\left(w_{[M,N_0]}^\theta\right)^t y}{\|w_{[M,N_0]}^\theta\|} \right) - \theta_\gamma^x \end{aligned} \quad (55)$$

$$\begin{aligned} &\leq \left(\theta_\gamma^x \times \frac{\left(w_{[M,N_0]}^\theta\right)^t y}{\|w_{[M,N_0]}^\theta\|} \right) - \theta_\gamma^x \\ &\leq 0, \end{aligned} \quad (56)$$

where ineq. (56) follows by the assumption that $\left(w_{[M,N_0]}^\theta\right)^t y \leq \theta$ and ineq. (57) follows since $\theta_\gamma^x \geq 0$ and since, by assumption, $\left(w_{[M,N_0]}^\theta\right)^t y \leq \theta = 0$. It follows that $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}(y) > 0$. Similarly, $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}(y) \leq 0$ whenever $\left(w_{[M,N_0]}^\theta\right)^t y > \theta$. All other cases result in y being non-anomalous. ■

Corollary 64: $\hat{\theta}_\gamma^x = -\theta_\gamma^x$ and $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x} = \phi_{[M,N_0]}^{-\theta_\gamma^x}$.

Proof: Follows directly from thm. (63) along with eqs. (52) and (53). ■

Remark 65: The shifted regression hyperplane, $\phi_{[M,N_0]}^{\theta^x}(y) = 0$, and its reflection across $\left(w_{[M,N_0]}^\theta\right)^t y = \theta$ given by $\hat{\phi}_{[M,N_0]}^{\hat{\theta}_\gamma^x}(y) = 0$, combine to segregate anomalous data points from non-anomalous data points.

VI. THE SUPPORT VECTOR DETECTOR

The contrapositive of thm. (62) requires that the absence of linear separability in the projected space results in the set of anomalies being empty by necessity. Hence, $\Omega = X_{[M,N_0]}^c$ so that all of the data belongs to one class. Furthermore, by calculating the distance from points in $X_{[M,N_0]}$ to the regression hyperplane, thms. (62) and (63) provide the means for identifying the class boundary to be used when determining if certain data points are anomalous, with ineq. (50) or (51) defining the shifted regression hyperplane. By identifying the set of data points in $X_{[M,N_0]}$ which minimize the distance to each section of the regression hyperplane, the set of support vectors defining half of the bound of the intermediate region is found. Similarly, by identifying the set of data points in $X_{[M,N_0]}^c$ which minimize the distance to the half of the bound which was previously found and by maximizing the distance to each section of the regression hyperplane, the set of support vectors defining the other half of the bound of the intermediate region is found.

Theorem 66: If $X_{[M,N_0]} \neq \emptyset$, then $X_{[M,N_0]}$ contains the set of all support vectors which bound half of the intermediate region.

Proof: Iteratively, apply thm. (62) (at most) N_0^2 times to find each $x \in X_{[M,N_0]}$ which satisfies eq. (47) by using uniformity to successively remove class C_h , where $h \in \{1, \dots, N_0\}$ gives rise to $d(x, H_{[M,N_0]}^\theta) = d_{[M,N_0]}^\theta$, while also removing all $y \in X_{[M,N_0]}$ such that $d(y, H_{[M,N_0]}^{\theta_\gamma})$ is minimized, for some fixed $\gamma \in (0, 1)$. The resulting set remains uniformly distributed in the projected space. Repeat this process (at most) N_0^2 times to find a set $V_{[M,N_0]}$. Since each $x \in V_{[M,N_0]} \subset X_{[M,N_0]}$ satisfies eq. (47), then by construction, $V_{[M,N_0]}$ contains the set of support vectors which bound half of the intermediate region. ■

Corollary 67: $X_{[M,N_0]}^c$ contains the set of all support vectors which bound the other half of the intermediate region.

Proof: Let $Y_{[M,N_0]} = X_{[M,N_0]}^c$ and apply thm. (66). ■

Remark 68: By necessity, each of the support vectors contained in $X_{[M,N_0]}^c$ are those data points closest in distance to the boundary of C_h , $h \in \{1, \dots, N_0\}$ when projected onto the vector normal to $H_{[M,N_0]}^{\theta_\gamma}$ which passes through $x_{(h)}$, while the support vectors contained within $X_{[M,N_0]}$ are found by minimizing the distance to $H_{[M,N_0]}^{\theta_\gamma}$ from the set of anomalies which are exposed as a result of $R(M, N)$ shrinking to $R(M, N_0)$.

Definition 69: Suppose $x \in X_{[M,N_0]}$ satisfies eq. (47). The *equivalency class of x* (denoted $[x]$) is the set of all vectors $y \in \mathcal{X}_n$ such that either y satisfies eq. (47) or

$$d(y, H_{[M,N_0]}^{\theta_\gamma}) \leq d_{[M,N_0]}^\theta.$$

Theorem 70: (Support Vector Characterization) If $X_{[M,N_0]} \neq \emptyset$ and $x \in X_{[M,N_0]}$ satisfies eq. (47), then $[x] \neq \emptyset$ and $[x]$ defines the complete set of support vectors for the hyperplanes of maximally-separating distance between $X_{[M,N_0]}$ and $X_{[M,N_0]}^c$.

Proof: Follows directly from thms. (62, 63, 66), cor. (64) and def. (69). ■

VII. CONCLUSIONS

It was shown that by mapping (possibly) higher dimensional data into a partitioned 2-dimensional space, a critical radius of connectivity could be found such that when radii are less than the critical value, then clusters of data points form. Furthermore, with the critical value defined as a function of the number of data points and the expected number of classes to form, the union of sets of regularized data points could be linearly segregated away from all other data points and that the normal vector to the shifted linear hyperplane defines weights of a 2-class discriminating neural network, which determines the support vectors of the formed binary classes. As such, the process of separating regularized data from all other (anomalous) data amounts to finding a critical radius, which is estimated in section (IV). After formation of the regularized classes by using the critical radius as the upper bound on the distance measure, we are able to formulate three estimates of the distribution of the regularized data.

The regression hyperplane of the union amounts to a global, stationary estimate of the non-stationary mixture. Furthermore, by formulating a regression hyperplane for each individual class of the mixture distribution, we can estimate the non-stationary distribution of the union as a sequence of piecewise linear (stationary) estimates, one for each class. Lastly, a piecewise, non-linear (non-stationary) estimate of the mixture is determined by the mode of each class.

Finally, by finding the $x \in X_{[M,N_0]}$ which provides the necessary shift for some $\gamma \in (0, 1)$, the task of segregating the union of regularized classes from the set of anomalies is completed, without the need of numerical techniques for estimation of a synaptic weight vector, since the original regression hyperplane $H_{[M,N_0]}^\theta$ is an analytic least squares solution, as determined by the union of the regularized classes. As the neural network detects the complete set of support vectors providing for a maximal-distance region between the union of regularized classes and the set of anomalies, no techniques such as Lagrange multipliers are required in order to determine the set of support vectors.

APPENDIX

A. Graph

Proposition 71: If $r < r'$, then $G(\mathcal{X}_n; r) \subseteq G(\mathcal{X}_n; r')$.

Proof: Suppose $r < r'$. If $\langle x, y \rangle_r \in G(\mathcal{X}_n; r)$, then $d(x, y) \leq r < r'$ so that $\langle x, y \rangle_{r'} \in G(\mathcal{X}_n; r')$. Hence, $G(\mathcal{X}_n; r) \subseteq G(\mathcal{X}_n; r')$. ■

B. Increasing Property

Lemma 72: $|\mathcal{A}_{[n,\rho]}^r| \leq 1$.

Proof: If $\mathcal{A}_{[n,\rho]}^r = \emptyset$, then there is nothing to prove. Thus, suppose that $\mathcal{A}_{[n,\rho]}^r$ occurs and $\langle C \rangle_r \in \mathcal{A}_{[n,\rho]}^r$. Since $\rho_n(C) \geq \rho > 1/2$, then all other connected components are of order strictly less than half of all points. Therefore, $|\mathcal{A}_{[n,\rho]}^r| = 1$. ■

Proposition 73: $\mathcal{A}_{[n,\rho]}^r$ is an increasing property in r .

Proof: Suppose $\langle C \rangle_r \in \mathcal{A}_{[n,\rho]}^r$ and fix arbitrary $r' > r$. Then, $d(x, y) \leq r < r'$ for all $x, y \in \langle C \rangle_r$. Thus, $\langle C \rangle_r \subseteq \langle C \rangle_{r'}$ implies $N = |\langle C \rangle_r| \leq |\langle C \rangle_{r'}|$. Hence, $\langle C \rangle_r \in \mathcal{A}_{[n,\rho]}^r$ implies $\langle C \rangle_{r'} \in \mathcal{A}_{[n,\rho]}^r$. Since $r' > r$ is arbitrary, then $\mathcal{A}_{[n,\rho]}^r$ is an increasing property in r . ■

Proposition 74: $\mathcal{A}_{[n,\rho]}^r$ is a decreasing property in n .

Proof: Suppose $\langle C \rangle_r \in \mathcal{A}_{[n,\rho]}^r$. If $n' < n$, then $|\langle C \rangle_r|/n' > |\langle C \rangle_r|/n \geq \rho$ so that $\langle C \rangle_r \in \mathcal{A}_{[n',\rho]}^r$. Hence, $\mathcal{A}_{[n,\rho]}^r \subseteq \mathcal{A}_{[n',\rho]}^r$. Since $n' < n$, then $\mathcal{A}_{[n,\rho]}^r$ is decreasing in n . ■

C. Probability Measure

Proposition 75: The property $\mathcal{A}_{[n,\rho]}^r$ is P -measurable.

Proof: For $x, y \in \mathcal{X}_n$ and $S \subseteq \mathcal{X}_n$, define the state on $\langle x, y \rangle_r$ to be 1 if and only if $\langle x, y \rangle_r \in G(S; r)$ and -1 otherwise. Then, S mutually determines an element $\omega_S \in \Omega = \{-1, 1\}^{\mathcal{X}_n}$ so that S is P -measureable. Since $\mathcal{A}_{[n,\rho]}^r$ is

the property that there exists $\omega_S \in \Omega$ mutually determined by $S \subseteq \mathcal{X}_n$ such that $(\max_{y \in S} | < C_y >_r |)/n \geq \rho$, then $\mathcal{A}_{[n,\rho]}^r$ is P -measureable. ■

Proposition 76: $P(\mathcal{A}_{[n,\rho]}^r)$ is a non-decreasing function of r .

Proof: Suppose $r_1 \leq r_2$. Since $\mathcal{A}_{[n,\rho]}^r$ is an increasing property in r by prop. (73), then $\mathcal{A}_{[n,\rho]}^{r_1} \subseteq \mathcal{A}_{[n,\rho]}^{r_2}$ so that $P(\mathcal{A}_{[n,\rho]}^{r_1}) \leq P(\mathcal{A}_{[n,\rho]}^{r_2})$ by properties of probability measures. Thus, $P(\mathcal{A}_{[n,\rho]}^r)$ is non-decreasing in r . ■

Proposition 77: $P(\mathcal{A}_{[n,\rho]}^r)$ is a non-increasing function of n .

Proof: Suppose $n' < n$. Since $\mathcal{A}_{[n,\rho]}^r$ is a decreasing property in n by prop. (74), then $\mathcal{A}_{[n,\rho]}^r \subseteq \mathcal{A}_{[n',\rho]}^r$ so that $P(\mathcal{A}_{[n,\rho]}^r) \leq P(\mathcal{A}_{[n',\rho]}^r)$ by properties of probability measures. Thus, $P(\mathcal{A}_{[n,\rho]}^r)$ is non-increasing in n . ■

D. Connection Radius

Proposition 78: $r(n, \rho, \epsilon)$ is a non-decreasing function of ϵ .

Proof: Suppose $\epsilon_1, \epsilon_2 \in (0, \frac{1}{2})$ such that $\epsilon_1 \leq \epsilon_2$. Define $r_1 = r(n, \rho, \epsilon_1)$ and $r_2 = r(n, \rho, \epsilon_2)$ and suppose $r_1 > r_2$. Since $P(\mathcal{A}_{[n,\rho]}^r)$ is non-decreasing in r by prop. (76), then $P(\mathcal{A}_{[n,\rho]}^{r_1}) \geq P(\mathcal{A}_{[n,\rho]}^{r_2}) \geq \epsilon_2 \geq \epsilon_1$. Hence, $r_2 \in \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_1\}$ and $r_2 < r_1 = \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_1\}$. Contradiction. Thus, $r_1 \leq r_2$ so that $r(n, \rho, \epsilon)$ is non-decreasing in ϵ . ■

Lemma 79: If $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$, then $\mathcal{X}_n = \{x \in \mathcal{X}_n : d(x, y) \leq R\}$ for all fixed $y \in \mathcal{X}_n$.

Proof: Clearly, $\{x \in \mathcal{X}_n : d(x, y) \leq R\} \subseteq \mathcal{X}_n$. Conversely, fix any $y \in \mathcal{X}_n$. For every $x \in \mathcal{X}_n$, it is true that $d(x, y) \leq 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\} = R$. Hence, $\mathcal{X}_n \subseteq \{x \in \mathcal{X}_n : d(x, y) \leq R\}$ for all fixed $y \in \mathcal{X}_n$. Thus, $\mathcal{X}_n = \{x \in \mathcal{X}_n : d(x, y) \leq R\}$ for all fixed $y \in \mathcal{X}_n$. ■

Corollary 80: If $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$, then $< C_y >_R \in \mathcal{A}_{[n,\rho]}^R$ for all $y \in \mathcal{X}_n$ and $n \geq 1$.

Proof: Fix an arbitrary $y \in \mathcal{X}_n$. By lemma (79), if $< C_y >_R = \{x \in \mathcal{X}_n : d(x, y) \leq R\}$, then $< C_y >_R = \mathcal{X}_n$ so that $| < C_y >_R | = |\mathcal{X}_n| = n$. Therefore, since $y \in \mathcal{X}_n$ is arbitrary, then $< C_y >_R \in \mathcal{A}_{[n,\rho]}^R$ for all $y \in \mathcal{X}_n$ and $n \geq 1$. ■

Corollary 81: If $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$, then $P(\mathcal{A}_{[n,\rho]}^R) = 1$ for all $n \geq 1$.

Proof: By lemma (79) and cor. (80), it is true that $\mathcal{X}_n \in \mathcal{A}_{[n,\rho]}^R$ for all $n \geq 1$ and $\rho \in (\frac{1}{2}, 1)$. Thus, $\mathcal{A}_{[n,\rho]}^R \neq \emptyset$ for all $n \geq 1$ and $\rho \in (\frac{1}{2}, 1)$. Hence, $P(\mathcal{A}_{[n,\rho]}^R) = 1$ for all $n \geq 1$. ■

Lemma 82: If $R = 2 * \max\{d(x, y) : x, y \in \mathcal{X}_n\}$, then $0 < r(n, \rho, \epsilon) \leq R$ for all $\epsilon \in (0, \frac{1}{2})$.

Proof: By lemma (79), it is true that $\mathcal{X}_n = \{x \in \mathcal{X}_n : d(x, y) \leq R\}$ for all fixed $y \in \mathcal{X}_n$. Therefore, $P(\mathcal{A}_{[n,\rho]}^R) = 1 \geq \epsilon$ for all $\epsilon \in (0, \frac{1}{2})$. Suppose that $\epsilon_0 \in (0, \frac{1}{2})$ exists such that $r_0 = r(n, \rho, \epsilon_0) > R$. Thus, $\mathcal{A}_{[n,\rho]}^{r_0} \subseteq \mathcal{A}_{[n,\rho]}^R$ so that

$$1 = P(\mathcal{A}_{[n,\rho]}^R) \leq P(\mathcal{A}_{[n,\rho]}^{r_0})$$

since $P(\mathcal{A}_{[n,\rho]}^r)$ is non-increasing in n by prop. (80), non-decreasing in r by prop. (76) and by properties of probability measures. Hence, $P(\mathcal{A}_{[n,\rho]}^{r_0}) = 1$. But, then $R \in \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_0\}$ and $R < r_0 = \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_0\}$. Contradiction. Thus, $0 < r_0 \leq R$. Therefore, $0 < r(n, \rho, \epsilon) \leq R$ for all $\epsilon \in (0, \frac{1}{2})$. ■

Proposition 83: Suppose $\{\epsilon_k \in (0, \frac{1}{2})\}_{k \geq 1}$ is any convergent sequence such that $\epsilon_k \rightarrow \epsilon_0$. Define $r_k = r(n, \rho, \epsilon_k)$ and $r_0 = r(n, \rho, \epsilon_0)$. For arbitrary $\xi > 0$, if $\{k \geq 1 : |P(\mathcal{A}_{[n,\rho]}^{r_k}) - P(\mathcal{A}_{[n,\rho]}^{r_0})| \geq \xi\}$ is a set of measure zero, then $r_k \rightarrow r_0$ as $k \rightarrow \infty$.

Proof: If $\xi > 0$ is arbitrary and $\{k \geq 1 : |P(\mathcal{A}_{[n,\rho]}^{r_k}) - P(\mathcal{A}_{[n,\rho]}^{r_0})| \geq \xi\}$ is a set of measure zero, then

$$P(\mathcal{A}_{[n,\rho]}^{r_k}) = P(\mathcal{A}_{[n,\rho]}^{r_0}) \geq \epsilon_0$$

for all $k \geq 1$. Hence, $r_k \in \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_0\}$ for all $k \geq 1$. Thus,

$$\begin{aligned} \lim_{k \rightarrow \infty} r_k &= \lim_{k \rightarrow \infty} r(n, \rho, \epsilon_k) \\ &= \lim_{k \rightarrow \infty} \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_k\} \end{aligned} \quad (58)$$

$$\begin{aligned} &= \inf\{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_0\} \\ &= r(n, \rho, \epsilon_0) \\ &= r_0 \end{aligned} \quad (59)$$

where eq. (58) and eq. (59) follow since $r_k \in \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_k\} \cap \{r > 0 : P(\mathcal{A}_{[n,\rho]}^r) \geq \epsilon_0\}$ for all $k \geq 1$ and $\epsilon_k \rightarrow \epsilon_0$ as $k \rightarrow \infty$. ■

REFERENCES

- [1] Alpaydin, E. (2010), *Introduction to Machine Learning, Second Edition*, The MIT Press.
- [2] Bai, X.; Kumar, S.; Xuan, D.; Yun, Z.; Lai, T. (2006), "Deploying Wireless Sensors to Achieve Both Coverage and Connectivity", *ACM MobiHoc'06, May 22 - 25, 2006, Florence, Italy*
- [3] Bai, H.; Chen, X.; Guan, X. (2006), "Preserving Coverage for Wireless Sensor Networks of Nodes with Various Sensing Ranges", *Proceedings of 2006 IEEE International Conference on Networking, Sensing and Control, ICNSC 2006, Ft. Lauderdale, Florida, USA, April 23-25, 2006, 5459*.
- [4] Balister, A.; Bollobas, B.; Sarkar, A.; Walters, M. (2005), "Connectivity of Random k-Nearest Neighbour Graphs", *Advances in Applied Probability, Vol. 37, no. 1, pp. 1 - 24, 2005*.
- [5] Balister, A.; Bollobas, B.; Sarkar, A.; Kumar, S. (2007), "Reliable Density Estimates for Coverage and Connectivity in Thin Strips of Finite Length", *ACM MobiCom'07, Sep. 9 - 14, 2007*
- [6] Barnes, R.; Burkett, T. (2010), *Structural Redundancy and Multiplicity in Corporate Networks, International Network for Social Network Analysis (INSNA), Volume 30, Issue 2, pp. 4 - 20, 2010*
- [7] Benes, V.; Rataj, J. (2004), *Stochastic Geometry: Selected Topics*, Kluwer Academic Publishers.
- [8] Bettstetter, C. (2002), "On the Minimum Node Degree and Connectivity of a Wireless Multihop Network", *ACM MobiHoc'02*
- [9] Bhondekar, A.P.; Vig, R.; Singla, M.L.; Ghanshyam, C.; Kapur, P. (2009), "Genetic Algorithm Based Node Placement Methodology for Wireless Sensor Networks", *Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, IMECS 2009, Mar. 18 - 20, 2009, Hong Kong*
- [10] Blanchard, G.; Bousquet, O.; Massart, P. (2008), *Statistical Performance of Support Vector Machines, The Annals of Statistics, Volume 36, No. 2, pp. 489 - 531, 2008*

- [11] Bollobas, B. (1991), "Random Walks", *Proceedings of Symposia in Applied Mathematics*, Vol. 9, 1991, pp. 1 - 20
- [12] Bourgain, J.; Kahn, J.; Kalai, G.; Katznelson, Y.; Linial, N. (1992), "The Influence of Variables in Product Spaces", *Israel Journal of Mathematics*, Vol. 77, No. 1 - 2, 1992, pp. 55 - 64
- [13] Cai, Haiyan; Jia, Xiaohua; Mo, Sha (2010), "Critical Sensor Density for Partial Connectivity in Large Area Wireless Sensor Networks", *Proceedings IEEE InfoCom'10*, 2010
- [14] Carroll, D.E.; Goel, A. (2004), *Lower Bounds for Embedding into Distributions over Excluded Minor Graph Families*, *Lecture Notes in Computer Science*, Volume 3221, pp. 146-156, 2004
- [15] Chandola, V.; Banerjee, A.; Kumar, V. (2009), *Anomaly Detection: A Survey*, *ACM Computing Surveys*, Volume 41, Issue 3, Article 15, 2009
- [16] Chen, Y.; Chuah, C.; Zhao, Q. (2005), "Sensor Placement for Maximizing Per Unit Cost in Wireless Sensor Networks", *Proc. IEEE Military Communications Conf.*, Oct. 2005
- [17] Chung, K.M.; Cao, W.C.; Sun, C.L.; Lin, C.J. (2003), *Decomposition Methods for Linear Support Vector Machines*, *Acoustics, Speech, and Signal Processing*, 2003. *Proceedings. (ICASSP '03)*. 2003 *IEEE International Conference on (Volume:4)*, Volume 4, pp. 868 - 871, 2003
- [18] Coffman Jr., E.G.; Shor P.W. (2005), "A Simple Proof of the $O(\sqrt{n} \log^{3/4}(n))$ Upright Matching Bound", *SIAM J. Disc Math*, Vol. 4, No. 1, pp. 48 - 57, 2005.
- [19] Consul, P.C. (1989), *Generalized Poisson Distributions Properties and Applications*, Marcel Dekker.
- [20] Csaki, E. (1997), "Some Results for Two Dimensional Random Walks", *Advances in Combinatorial Methods and Applications to Probability and Statistics*, 1997
- [21] Davidson, I.; Ward, M. (2001), *A Particle Visualization Framework for Clustering and Anomaly Detection*, *ACM KDD Workshop on Visual Data Mining*, 2001
- [22] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977), *Maximum Likelihood for Incomplete Data via the EM Algorithm*, *Journal of the Royal Statistical Society, Series B (Methodological)*, Volume 39, No. 1., pp. 1 - 38, 1977
- [23] Durrett, R. (1991), *The TEX Probability : Theory and Examples.*, Duxbury.
- [24] Efrat, A.; Itai, A.; Katz, M. (1996), "Improvements on Bottleneck Matching and Related Problems Using Geometry", *Proceedings of the 12th Annual Symposium on Computational Geometry*, pp. 301 - 310, 1996
- [25] Franceschetti, M.; Booth, L.; Cook, M.; Meester, R.; Bruck, J. (2005), "Continuum Percolation with Unreliable and Spread-Out Connections", *Journal of Statistical Physics*, Vol. 118, No. 3/4, Feb. 2005.
- [26] Friedgut, E.; Kalai, G. (1996), "Every Monotone Graph Property has a Sharp Threshold", *Proceedings of the American Mathematical Society*, Vol. 124, No. 10, 1996, pp. 2993 - 3002
- [27] Geman, S.; Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 6 (Nov. 1984)
- [28] Georgii, Hans-Otto (1988), *Gibbs Measures and Phase Transitions*, Walter de Gruyter.
- [29] Gilbert, E.N. (1961), "Random Plane Networks", *J. Soc. Indust. Appl. Math.*, Vol. 9, No. 4, Dec. 1961, pp. 533 - 543
- [30] Glauche, I.; Krause, W.; Sollacher, R.; Greiner, M. (2003), "Continuum Percolation of Wireless Ad Hoc Communication Networks", *Physica A*, 325, 2003, pp. 577 - 600
- [31] Goel, A.; Rai, S.; Krishnamachari, B. (2005), "Monotone Properties of Random Geometric Graphs have Sharp Thresholds", *The Annals of Applied Probability*, Vol. 15, No. 4 (Nov. 2005), pp. 2535 - 2552
- [32] Grimmett, Geoffrey (1999), *Percolation*, Springer-Verlag.
- [33] Grimmett, Geoffrey (2006), *The Random-Cluster Model*, Springer-Verlag.
- [34] Guthrie, D.; Guthrie, L.; Allison, B.; Wilks, Y. (2007), *Unsupervised Anomaly Detection*, *IJCAI-07*, pp. 1624 - 1628, 2007
- [35] Guyon, Xavier. (1995), *Random Fields on a Network: Modeling, Statistics and Applications*, Springer.
- [36] Han, X.; Cao, X.; Zhang, Y.; Lloyd, E.; Shen, C. (2007), "Deploying Directional Wireless Sensors with Guaranteed Coverage and Connectivity", *IEEE InfoCom 2007*
- [37] Haykin, S. (1994), *Neural Networks, A Comprehensive Foundation*, Macmillan College Publishing Company, Inc.
- [38] Hogg, R.V.; McKean, J.W.; Craig, A.T. (2005), *Introduction to Mathematical Statistics*, Pearson Prentice Hall.
- [39] Hou, T.; Li, V. (1986), "Transmission Range Control in Multi-hop Packet Radio Networks", *IEEE Trans. on Communications*, Vol. 34, No. 1, 1986, pp. 38 - 44
- [40] Hsu, C.W.; Lin, C.J. (2002), *A Simple Decomposition Method for Support Vector Machines*, *Machine Learning*, Volume 46, pp. 291 - 314, 2002
- [41] Huang, L.; Nguyen, X.L.; Garofalakis, M.; Jordan, M.; Joseph, A.D.; Taft, N. (2006), *In-Network PCA and Anomaly Detection*, *NIPS*, 2006
- [42] Joachims, T. (1998), *Making Large-Scale Support Vector Machine Learning Practical*, *Advances in Kernel Methods: Support Vector Machines*, Scholkopf, B., Burges, C., Smola, A., Eds. Cambridge, MA: MIT Press 1998
- [43] Kar, A. (2003), *Weyl's Equidistribution Theorem*, *Resonance*, pp. 30 - 37, 2003
- [44] Keerthi, S.S.; Lin, C.J. (2003), *Asymptotic Support Vector Machines with Gaussian Kernel*, *Neural Computation*, Volume 15, Issue 7, pp. 1667 - 1689, 2003
- [45] Keerthi, S.S.; Shevade, S.K.; Battacharyya, C.; Murthy, K.R.K. (2001), *Improvements to Platt's SMO Algorithm for SVM Classifier Design*, *Neural Computation*, Volume 13, pp. 637 - 649, 2001
- [46] Kirilyuk, K.P. (2004), "Complex Dynamics of Autonomous Communication Networks and the Intelligent Communication Paradigm", *Report at the International Workshop on Autonomic Communication*, Berlin, 1819 October 2004.
- [47] Kleinrock, L.; Silvester, J. (1978), "Optimal Transmission Radii for Packet Radio Networks or Why Six is a Magic Number", *NTC'78*
- [48] Kolmogorov, A.N.; Fomin, S.V. (1970), *Introductory Real Analysis*, Dover Publications.
- [49] Kuperberg, W. (1989), "Covering the Plane with Congruent Copies of a Convex Body", *Bulletin of the London Mathematical Society*, Vol. 21, pp. 82 - 86, 1989
- [50] Lee, W.; Xiang, D. (2001), *Information-Theoretic Anomaly Detection*, *Proceedings. 2001 IEEE Symposium on Security and Privacy*, pp. 130 - 143, 2001
- [51] Leighton, T.; Shor, P. (1989), "Tight Bounds for Minimax Grid Matching with Applications to the Average Case Analysis of Algorithms", *Combinatorica*, Vol. 9, No. 2, 1989, pp. 161 - 187
- [52] Lin, C.J. (2001), *On the Convergence of the Decomposition Method for Support Vector Machines*, *IEEE Transactions on Neural Networks*, Volume 12, Issue 6, pp. 1288 - 1298, 2001
- [53] Liu, J.; Adler, M.; Towsley, D.; Yun, Z.; Zhang, C. (2006), "On Optimal Communication Cost for Gathering Correlated Data through Wireless Sensor Networks", *ACM MobiCom'06*, Sep. 24 - 29, 2006
- [54] Mamelì, V.; Musio, M. (2013), *A Generalization of the Skew-Normal Distribution: The Beta Skew-Normal*, *Communications in Statistics - Theory and Methods*, Volume 42, pp. 2229-2244, 2013
- [55] Mangasarian, O.L.; Musicant, D.R. (2000), *Active Set Support Vector Machine Classification*, *Advances in Neural Information Processing Systems*, pp. 577-583, 2000
- [56] Maselli, G.; Deri, L.; Suin, S. (2003), *Design and Implementation of an Anomaly Detection System: An Empirical Approach*, *Proceedings of Terana Networking Conference*, Zagreb Croatia, 2003
- [57] Meester, Ronald; Roy, Rahul (1996), *Continuum Percolation*, Cambridge University Press.
- [58] Miller, W.T.; Sutton, R.S.; Werbos, P.J. (1990), *Neural Networks for Control*, The MIT Press.
- [59] Moise, E. (1990), *Elementary Geometry from an Advanced Standpoint*, Addison Wesley.

- [60] Murphy, Robert (2011), *Partial Connectivity in Wireless Sensor Networks with Applications*, UMI Proquest.
- [61] Murphy, R. (2015), *Estimating the Mean Number of K-Means Clusters to Form*, *arXiv*, ID 1503.03488, 2015
- [62] Osuna, E.; Freund, R.; Girosi, F. (1997), *Improved Training Algorithm for Support Vector Machines*, *Proc. IEEE NNSP '97*, 1997
- [63] Platt, J.C. (1998), *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, *Advances in Kernel Methods: Support Vector Machines*, Scholkopf, B., Burges, C., Smola, A., Eds. Cambridge, MA: MIT Press 1998
- [64] Rai, A. (2004), "The Spectrum of a Random Geometric Graph is Concentrated", <http://arxiv.org/PS/cache/math/pdf/0408/0408103.pdf>, Sep. 2004
- [65] Saligrama, V.; Zhao, M. (2012), *Local Anomaly Detection*, *Journal of Machine Learning Research*, W&CP, Volume 22, pp. 969 - 983, 2012
- [66] Santi, P.; Blough, D. (2003), "The Critical Transmitting Range for Connectivity in Sparse Wireless Ad-Hoc Networks", *IEEE Trans. on Mobile Computing*, Vol. 2, No. 1, 2003, pp. 25 - 39
- [67] Schechter, E. (1996), *Handbook of Analysis and Its Foundations.*, Academic Press.
- [68] Shiryaev, A.N. (1996), *Probability*, Springer Verlag.
- [69] Shor, P.W.; Yukich, J.E. (1991), "Minimax Grid Matching and Empirical Measures", *The Annals of Probability*, Vol. 19, No. 3 (1991), pp. 1338 - 1348
- [70] Song, X.; Wu, M.; Jermaine, C.; Ranka, S. (2007), *Conditional Anomaly Detection*, *IEEE Transactions on Knowledge and Data Engineering*, Volume 19, Issue 5, pp. 631 - 635, 2007
- [71] Steinwart, I.; Hush, D.; Scovel, C. (2005), *A Classification Framework for Anomaly Detection*, *Journal of Machine Learning Research*, Volume 6, pp. 211 - 232, 2005
- [72] Takagi, H.; Kleinrock, L. (1984), "Optimal Transmission Ranges for Randomly Distributed Packet Radio Terminals", *IEEE Trans. on Communications*, Vol. 32, No. 3, 1984, pp. 246 - 257
- [73] Takahashi, N.; Nishi, T. (2006), *Global Convergence of Decomposition Learning Methods for Support Vector Machines*, *IEEE Transactions on Neural Networks*, Volume 17, Issue 6, pp. 1362 - 1369, Nov. 2006
- [74] Takahashi, N.; Jun, G.; Nishi, T. (2008), *Global Convergence of SMO Algorithm for Support Vector Regression*, *IEEE Transactions on Neural Networks*, Volume 19, Issue 6, pp. 971 - 982, Mar. 2008
- [75] Vapnik, V. (1982), *Estimation of Dependencies Based Upon Empirical Data*, Springer-Verlag.
- [76] Wan, P.; Yi, C. (2004), "Asymptotic Critical Transmission Radius and Critical Neighbor Number for k-Connectivity in Wireless Ad-Hoc Networks", *ACM MobiHoc'04*
- [77] Wang, X.; Guoliang, X.; Zhang, Y.; Lu, C.; Pless, R.; Gill, C. (2003), "Integrated Coverage and Connectivity Configuration in Wireless Sensor Networks", *SenSys '03*, Nov. 5 - 7, 2003, Los Angeles, CA, USA
- [78] Xing, G.; Wang, X.; Zhang, Y.; Lu, C.; Pless, R.; Gill, C. (2005), "Integrated Coverage and Connectivity Configuration for Energy Conservation in Sensor Networks", *ACM Trans. on Sensor Networks*, Vol. 1, No. 1, 2005, pp. 36 - 72
- [79] Xue, F.; Kumar, P.R. (2006), "On the θ -Coverage and Connectivity of Large Random Networks", *Joint Special Issue of the IEEE Trans. on Information Theory and the IEEE/ACM Trans. on Networking on "Networking and Information Theory"*, (May 2006)
- [80] Yu, Y.; Hong, B.; Prasanna, V.K. (2005), "Communication Models for Algorithm Design Automation in Wireless Sensor Networks", *Proceedings of the IEEE International Parallel and Distributed Processing Symposium*, 2005
- [81] Zhang, H.; Hou, J. (2005), "Maintaining Sensing Coverage and Connectivity in Large Sensor Networks", *Ad Hoc & Sensor Wireless Networks*, Vol. 1, No. 1-2, 2005 pp. 89 - 124
- [82] Zhang, Y.; Chong, E.K.P.; Hannig, J.; Estep, D. (2010), "Continuum Limits of Markov Chains with Application to Network Modeling", *Proceedings of the 49th IEEE Conference on Decision and Control, CDC 2010, December 15-17, 2010*, pp. 6779 - 6784, 2010